EXPLORING THE ADDED VALUE OF A NUMBER LINE ASSESSMENT FOR KINDERGARTEN MATHEMATICS SCREENING

by

DAVID J. FURJANIC

A DISSERTATION

Presented to the Department of Special Education and Clinical Sciences
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

June 2021

DISSERTATION APPROVAL PAGE

Student: David J. Furjanic

Title: Exploring the Added Value of a Number Line Assessment for Kindergarten Mathematics Screening

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Special Education and Clinical Sciences by:

Ben Clarke	Chairperson and Advisor
Hank Fien	Core Member
Lillian Duran	Core Member
Joseph Nese	Core Member
Gerald Tindal	Institutional Representative

and

Kate Mondloch	Interim Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded June 2021.

DISSERTATION ABSTRACT

David J. Furjanic

Doctor of Philosophy

Department of Special Education and Clinical Sciences

June 2021

Title: Exploring the Added Value of a Number Line Assessment for
Kindergarten Mathematics Screening

Despite the importance of mathematical understanding for academic and occupational success, students in the United States are not meeting necessary levels of mathematics achievement. Multi-tiered systems of support (MTSS) provide a framework for schools to allocate resources to best support students. Universal screening, a key element of MTSS, employs brief assessments of critical academic skills to identify at-risk students. Despite advances in the screening for reading risk, research in mathematics screening is lacking. Current early numeracy screeners target number sense with mixed results. The mental number line is a potential construct for developing more advanced screening measures. The mental number line is a key developmental construct around which students organize their thinking and draw upon when working with elementary mathematics topics. The current study will explore the promise of using a number line assessment as part of a mathematics screening battery to identify students at risk.

CURRICULUM VITAE

NAME OF AUTHOR: David J. Furjanic

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
Millersville University of Pennsylvania, Millersville
The Pennsylvania State University, State College

DEGREES AWARDED:

Master of Science, Psychology, 2017, Millersville University of Pennsylvania
Bachelor of Science, Psychology, 2013, The Pennsylvania State University

AREAS OF SPECIAL INTEREST:

Multi-Tiered Systems of Supports in K-12 Education
Data-Based Decision-Making
Equity in School Practices

PROFESSIONAL EXPERIENCE:

Advanced Practicum Student, Center on Teaching and Learning, University of Oregon, Eugene, 2018 to 2020

Diversity and Retention Graduate Employee, Graduate School, University of Oregon, Eugene, 2019 to 2020

Teaching Assistant Graduate Employee, College of Education, University of Oregon, Eugene, 2018 to 2019

School Psychology Intern, Derry Township School District, Hershey, Pennsylvania, 2016 to 2017

Graduate Assistant, Psychology Department, Millersville University, Millersville, 2015 to 2016

Therapeutic Support Staff, Pennsylvania Counseling Services, Harrisburg, Pennsylvania, 2013 to 2014

PUBLICATIONS:

Clarke, B., Nelson, N., Kosty, D., Ketterlin-Geller, L., Smolkowski, K, Lesner, T., Furjanic, D., & Fien, H. (Under review). *Investigating the promise of a tier two sixth grade fractions intervention.*

Sutherland, M., Clarke, B., Nese, J. F. T., Strand Cary, M., Shanley, L., Furjanic, D., & Durán, L. (Submitted for review). *Investigating the utility of a kindergarten number line assessment compared to an early numeracy screening battery*.

ACKNOWLEDGEMENTS

I thank Dr. Ben Clarke for his assistance in the preparation of this manuscript. I would also like to thank Drs. Joseph Nese, Lillian Durán, Gerald Tindal, and Hank Fien for contributing their expertise as I refined this study. I am sincerely grateful for the Special Education and Clinical Sciences faculty, staff, and students as well as my family and loved ones for their boundless intellectual, emotional, and social support throughout my scholarship at the University of Oregon. The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grants R305K040081 and R305A080699 to the Center on Teaching and Learning at the University of Oregon and Pacific Institutes for Research.

TABLE OF CONTENTS

Chapter Page

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

I. INTRODUCTION

The State of Mathematics Achievement in the United States

Calls to improve mathematics achievement for our nation's students have been spurred by consistent and long-standing patterns of low achievement by students in the United States (National Mathematics Advisory Panel, 2008; National Research Council, 2001). The 2019 National Assessment for Educational Progress (NAEP) found that less than half of fourth grade students were proficient in mathematics. NAEP proficiency levels have remained relatively stable over the past decade and a half, demonstrating a protracted concern. Of even greater concern, average scores for historically disadvantaged students (such as students eligible for free/reduced lunch, attending urban schools, or identified with disabilities) had statistically significant drops from previous years (National Center for Education Statistics, 2015; National Center for Education Statistics, 2005-2019; OECD, 2012). Comparisons to international peers further illustrate the depth of the problem. Students in the United States are ranked in the lower half of students worldwide, with performance gaps increasing as students advance across grades (Olson, Martin, & Mullis, 2008). United States students will be hindered competing in both international and domestic job markets without secure mathematics skills as high-demand professions, including those in science, technology, engineering, mathematics (STEM), increasingly rely upon a strong mathematical foundation (National Science Board, 2015).

Mathematics difficulties reverberate beyond schooling and affect basic functional tasks for adults. Over half of adults cannot calculate a 10% tip for a meal and even more cannot calculate miles per gallon on a trip (Phillips, 2007). While a secure foundation in mathematics can afford opportunities, an insecure foundation in mathematics has educational, occupational, and functional life implications for our students (National Mathematics Advisory Panel, 2008).

American adults who struggle with mathematics are, by and large, products of the American school systems (Mcclure et al., 2017; Watts, Duncan, Siegler, & Davis-Kean, 2014). Numerical knowledge at age 7, or typically first grade, predicts socioeconomic status at age 42 even when controlling for IQ, reading achievement, and familial SES (Ritchie & Bates, 2013). Students enter school exhibiting individual differences and these differences compound over time resulting in expanding achievement gaps over time (Bodovski & Farkas, 2007; Jordan, Kaplan, & Hanich, 2002; Judge & Watson, 2011; Schulte & Stevens, 2015; Wei, Lenz, & Blackorby, 2013). Utilizing a large national dataset, the Early Childhood Longitudinal Study – Kindergarten Cohort (ECLS-K), Morgan and colleagues (2009) found that kindergarten students who entered and subsequently exited in the lowest 10th percentile in mathematics had a 70% chance of still being in the bottom 10th percentile in fifth grade. Given the cumulative nature of mathematics understanding, early foundational gaps in knowledge limit the acquisition of more advanced content (Duncan et al., 2007; Hiebert & Wearne, 1996; Judge & Watson, 2011).

The importance of serving at-risk students early is underscored by the fact that preventing academic difficulties saves significant time and resources compared to remediation approaches (Fletcher & Vaughn, 2009; Torgesen, 2000, 2002; Torgesen et al., 2001; Vaughn & Wanzek, 2014; Vaughn et al., 2011; Walker et al., 1996). Morgan and colleague's (2009) study found that of those students who entered kindergarten below the 10th percentile but exited kindergarten above the 10th percentile, only 30% were in the bottom 10th percentile in fifth grade. This phenomenon demonstrates the potential of early intervention during this time period to alter and promote favorable learning trajectories. The promise of early intervention has been codified within the reauthorization of IDEA (Individuals with Disabilities Education Act, 2004). Aligned with calls by the field (Gersten et al., 2009), IDEA (2004) emphasizes the prevention of protracted learning difficulties through early identification and intervention (Gersten et al., 2009).

Addressing Mathematics Needs Via Screening for Risk

Educators and schools can promote favorable trajectories for students. One avenue for promoting students' success is universally screening all students to identify who is most at risk for academic difficulties. Universal screening involves administering a brief assessment to all students in a school, typically at three timepoints throughout the year. Screening data is used to guide decisions on which individual students are at-risk and also to systematically gauge the health of the system as a whole (Albers & Kettler, 2014; Shinn, 2006; Simmons et al., 2000).

Screeners have the potential to supply schools with critical information for serving their students. Screening is a key feature in Multi-Tiered Systems of Support (MTSS) or Response to Intervention (RTI) frameworks. However, schools across the country utilize screening to varying degrees. Despite over 70% of schools reporting using an MTSS/RTI framework to support reading development, only 35% reported using this framework for mathematics (Balu et al., 2015). Successful MTSS implementation requires substantial resources, tools, training, and commitment on the part of a school or district (Fletcher & Vaughn, 2009; D. Fuchs & Fuchs, 2017). Critical to a successful MTSS framework are the measures around which universal and targeted decisions are made (Balu et al., 2015; D. Fuchs & Fuchs, 2017).

Early Mathematics Screening

The value in a screener hinges upon its ability to assess important constructs in a given content area. Early numeracy screeners most commonly assess aspects of number sense (Gersten et al., 2012). Number sense, while its definition varies across the field, is best described as a series of interrelated early mathematical competencies that serve as a foundation for the acquisition of more advanced concepts (Feigenson, Libertus, & Halberda, 2013; Griffin, Case, & Siegler, 1994; Jordan, Glutting, & Ramineni, 2010; Jordan, Kaplan, Olah, & Locuniak, 2006; Jordan, Kaplan, Ramineni, & Locuniak, 2009; Siegler & Lortie-Forgues, 2014; Siegler, Thompson, & Schneider, 2011; Starr, Libertus, & Brannon, 2013).

Number sense screeners are created with the developmental progression of number sense in mind. Young children exhibit the precursors to number sense prior to formal schooling. The development of number sense begins perceptually before children can visualize or cognitively represent and manipulate numbers. Initially, children need to engage with tangible quantities (such as balls, dots, or patterns). As they develop, they can visualize quantities and patterns with imagined objects. Once school-aged, children extend their ability to reason with numbers to greater quantities and transition from informal to formal number sense. This transition is, in part, aided by the introduction of and continued engagement with symbolic numbers. Students with well-developed number sense can fluently and accurately reason with, manipulate, and problem-solve with numbers and quantities in a base-ten system (Berch, 2005; Case et al., 1996; Gersten, Jordan, & Flojo, 2005).

Number sense screeners are intentionally created to capture students' proficiency at various points in mathematical development. Furthermore, number sense is an amalgamation of interrelated skills and the developmental progressions of these skills look different. Among the interrelated skills invoked in number sense are counting, magnitude comparison, number operations, and symbolic numerical understanding (Case, 1998; Clements, Sarama, & DiBiase, 2003; Cross, Woods, & Schweingruber, 2009).

The Components of Number Sense

Counting. Counting is an essential foundation to developing number sense (Hudson & Miller, 2005). As young as infancy, children exhibit the first signs of number sense through numerosity, or the beginning stages of understanding quantity (Gallistel & Gelman, 1992). Infants can perceptually subitize, or recognize small quantities without systematically counting (Clements, 1999; Starkey & Cooper, 1980; Wynn, Bloom, & Chiang, 2002), which includes the ability to discriminate that an array of 4 items is different than an array of 2 (Starkey, Spelke, & Gelman, 1990). By eighteen months, they exhibit greater understanding in being able to identify that the array of 4 is *greater than* the array of 2 (Cooper Jr, 1984).

Between the ages of two and three, children begin to learn the number words from one to ten. As children start to grapple with these numbers words and counting in parallel, they learn to associate each word with an object (Baroody, 2002; Wagner & Walters, 1982). In counting to a number, the child associates meaning to each item in the sequence as they touch each object once with an accompanying word (Cross et al., 2009;

Mix, Huttenlocher, & Levine, 2002). This skill of associating each object once and only once with a number while counting is called one-to-one correspondence. One-to-one correspondence sets the foundation for children to then learn cardinality, or the understanding that the last word said in the sequence holds meaning for the collection. In counting a group of five objects, “five” as the final word represents the set as a whole (Clements et al., 2003). Counting, as an aspect of number sense, manifests first as an ability to recognize items in a set and then gradually develops into an ability to attach specific meaning to quantities. As their list of known number words and numerals grows, children extend to greater quantities their ability to count using one-to-one correspondence in a fixed order and with cardinality (Clements et al., 2003; Cross et al., 2009).

Number Knowledge. Number knowledge, another skill invoked in number sense, first manifests when children begin to compare sets of objects. Before they understand numerical quantities, children often rely upon perceptual clues to decide which set is greater, such as which set is spaced farther apart (Cross et al., 2009). The well-known example of this behavior is children’s lack of understanding conservation. Imagine showing a child two sets of teddy bear counters each containing four teddy bears. The child may agree the two sets are equivalent if they look similar. Now imagine if one of the sets were adjusted so that its four teddy bears are spread out in a long line. A child who doesn’t understand conservation would likely assert that the long line of teddy bears is now greater than its twin because of the child’s reliance upon perceptual clues.

Once children become familiar with number words, they tend to rely upon these to make their judgments. A four year old, for example, would count each set of items and decide which is greater based on which number word was “farther” down the list or number line (Clements et al., 2003; Cross et al., 2009). Children in formal schooling increasingly extend their ability to reference their number word list to make judgements about magnitudes. Children in kindergarten leverage their budding knowledge of magnitude to conceptually subitize, or understand how larger quantities are composed of groups of smaller, more familiar quantities (Cross et al., 2009; Griffin, 2004; Jordan, Glutting, & Ramineni, 2010; Jordan et al., 2009).

By age 6, children integrate their knowledge of counting and magnitudes into a mental number line (Siegler & Booth, 2004). Referred to as a “central conceptual structure,” children’s construction of the mental number line is theorized to enable children to access the quantitative world in a way they could not previously (Griffin, 2002).

Number Operations. The beginnings of children’s ability to engage in number operations is also evident before school-aged years. Children as young as two exhibit preverbal mathematical numeration. They can recognize basic number operations of adding or taking away one object. For example, a toddler observing two balls being placed into a box and one being removed would expect one to remain (Wynn, 1992). Similarly, young children can recognize which set is greater when one item is added to only one of two equivalent groups. It is not until the age of five, however, that children can judge magnitudes for collections that did not begin equivalently (Clements et al., 2003; Cooper Jr, 1984). As children simultaneously develop their ability to count, they rely upon this skill to manipulate numbers. To add three to a set of four, children may initially count each set and then count the two sets together. Children advance into being able to count to four and directly continue counting three more times to reach seven. As children understand that the two addends are subsumed within the total, they can start counting at one of the quantities, such as four, and count on to the total, seven, from there (Clements et al., 2003).

Symbolic Number Understanding. Symbolic number understanding, such as recognizing the numeral “2” represents a set of two items, is also crucial to number sense. To secure this skill, students must be able to recognize the form of the numeral, produce the form accurately, and attach the correct meaning to the form. For example, a young child with secure number sense would be able to recognize a printed 6, reproduce the numeral, and understand that it represents a set of six items (Clements et al., 2003).

Screening for Number Sense

Number sense is comprised of key early numeracy skills that enable children to engage with more advanced mathematics (Gersten et al., 2005; Jordan et al., 2009). Due to the importance of early trajectories, early numeracy researchers have focused measure development on assessing components of this foundational construct. In a review of screeners, Gersten et al. (2012) make clear the prominence of number sense measures as a means to predict risk in mathematics. Researchers have leveraged observable tasks such as magnitude comparison or strategic counting in order to measure the construct of number sense (Chard et al., 2005; Conoyer, Foegen, & Lembke, 2016; Gersten et al., 2012; Lembke & Foegen, 2009; Mazzocco, 2005; Seethaler & Fuchs, 2010).

For example, VanDerHeyden et al. (2001) administered a set of three one-minute group-administered measures tapping into early components of number sense. Kindergarten students (n = 107) counted circles and wrote the numeral of the total, counted circles and selected the number of circles from a set, and drew a number of circles corresponding to a numeral. Using a smaller sample of students, the researchers explored validity of the measures in relation to retention at the end of the year. They found that the scores correctly predicted retention in 71.4% (or 5/7) of cases and promotion in 94.4% (or 17/18) of cases. Concurrent validity correlations ranged from .44 to .61.

The Number Knowledge Test (NKT; Okamoto & Case, 1996) was explored by Baker et al. (2002) and Gersten, Jordan and Flojo (2005) with a sample of more than 200 kindergarten students. The NKT is a 10-15 minute individually administered assessment of a student's procedural and conceptual knowledge of whole numbers. The NKT assesses components of number sense through increasingly complex counting, magnitude comparison, and number operation tasks. The researchers found that the NKT exhibited strong predictive validity to end-of-first grade outcomes on the SAT-9 Total Mathematics (r = .73).

Clarke and Shinn (2004) and Clarke, Baker, Smolkowski, and Chard (2008) assessed three components of number sense – number identification, quantity discrimination, and missing number – with kindergarten (n = 52) and first grade students (n = 111; with 1-10 and 1-20 target numbers, respectively). In the first task, students identified given numerals. With similar stimuli, students chose the greater quantity in a pair given two numerals for quantity discrimination. For missing number, students identified the number missing from a sequence of three consecutive units with the missing number in the first, middle, or last position (e.g. __, 4, 5 or 6, __, 8). Predictive validities were strong across both studies, ranging from .62 to .64 with standard achievement tests (Woodcock-Johnson Applied Problems subtest and the Stanford Early School Achievement Test, respectively).

In a 4-year longitudinal study, Mazzocco and Thompson (2005) followed 226 students from kindergarten to third grade to determine the best assessment or battery to predict mathematics difficulty. Measures included mathematics achievement, formal and informal mathematics ability, visual-spatial reasoning, and rapid automatized naming assessments. Four items within the battery best predicted mathematics difficulty (here defined as performance below the 10th percentile on a third-grade comprehensive mathematics measure). The items that best predicted math difficulty (reading numerals, number constancy, magnitude judgments, and mental addition of one-digit numbers) were all mathematics items and associated with components of number sense. Most importantly, these four items correctly classified 84% of third-grade students at-risk based upon their performance in kindergarten.

Seethaler and Fuchs (Seethaler & Fuchs, 2010) also explored screeners that assessed different components of number sense. The researchers administered a magnitude comparison measure and a multiple proficiency measure, Number Sense, to 196 kindergarten students in the fall and spring. At the end of first grade, they administered The Early Math Diagnostic Assessment and the KeyMath-Revised. Predictive validity of the fall screeners to the spring outcome measures ranged from .52 to .72. Classification accuracy was relatively high across both methods, ranging from .67 to .86.

Hampton et al. (2012) administered six measures tapping into number sense (counting, number identification, missing number, quantity discrimination, next number, and number facts) to kindergarten (*n* = 71) and first grade (*n* = 75) students weekly. The researchers found small to large predictive validities from fall to spring (ranging from .26 to .52) with the Broad Math Score of the Woodcock-Johnson Battery of Achievement-III (J. Cohen, 1992).

Across these seminal screening studies and others (L. S. Fuchs et al., 1994; Lee & Lembke, 2016; Lembke & Foegen, 2009), researchers have leveraged the construct of number sense to develop early numeracy screeners and to examine the relationship between current and future achievement in mathematics. Despite earnest efforts, research on early numeracy screeners has not produced optimal screening measures. Gersten et al.'s (2012) review found the median predictive validities for kindergarten students on magnitude comparison and strategic counting measures were "moderate" at .50 and .48, respectively (Gersten et al., 2012).

The Field's Approach to Screening

In search of refining screening practices, researchers have explored various structural approaches. Most mathematics screeners adopt a curriculum-sampling approach. Sampling from the curriculum tends to provide more information about specific domains within a grade rather than general mathematics proficiency (Foegen, Jiban, & Deno, 2007). Pulling from curricular objectives is useful for instructional decision-making and progress monitoring throughout a year but less so for screening in the fall (Vanderheyden, Codding, & Martin, 2017). Drawing upon skills which students have not yet been taught often invokes a floor effect, where a tool is unable to discriminate students along a spectrum because too many students scored within a narrow band close to the bottom. A floor effect hinders the tool’s ability to detect which students are at risk and/or whether a systemic problem exists (Vanderheyden et al., 2017).

An alternative to a single curriculum-sampled tool is a net of screeners that span multiple domains. This approach, in part necessitated by the nonlinear development of mathematics skills, tends to be more predictive of math performance than single-skill screeners (Gersten et al., 2012; Seethaler & Fuchs, 2010). For example, VanDerHeyden, Codding, and Martin (2017) found that a combined screening net of multi-skill computation, single skill computation, and concepts/applications tasks had high diagnostic accuracy for fourth and fifth grade students.

An elaborate net of multiple screening measures may more accurately determine students at risk, but each additional measure included in a screening battery costs significant amounts of instructional and personnel time. Educators must weigh the relative benefits of each measure in their screening battery against the value of the information it provides. Rather than comprehensively sampling across every mathematics domain for a given grade, established batteries of selected measures could be supplemented or replaced with an assessment of a central concept that integrates multiple skills.

The Promise of the Number Line

The mental number line is theorized to be a central concept around which students organize their mathematical thinking (Case et al., 1996; Laski & Siegler, 2007; Schneider et al., 2018; Siegler, 2016; Siegler et al., 2011). As children first grapple with numbers, they begin to place these quantities along a mental number line. The number words they learn take shape in a linear fashion as they understand each successive number is one greater than that before it (Clements et al., 2003). As they encounter larger quantities, their mental number line expands outwards to accommodate. At this stage, the number line may be more aptly called a number path, as students understand numbers only as integers, or whole numbers to jump to with each successive increment (Cross et al., 2009). As they work with fractions and decimals, their mental number line grows interstitially, becoming more detailed between quantities (Siegler, 2016) and as their understanding of rational numbers continues to expand the mental number line morphs from a series of connected, discrete integers to a continuous spectrum of potentially-infinite quantities.

Students draw upon this mental number line for various mathematical tasks (Schneider et al., 2018; Siegler & Lortie-Forgues, 2014; Siegler et al., 2011). When comparing magnitudes, for example, locating 11 and 14 on one's mental number line can enable a student to understand which has a greater magnitude (Siegler, 2016). Similarly, a student may tap into their mental number line to solve an addition problem such as "4 + 2" by referencing "4" on their mental number line and counting up two integers. Students may also draw upon this mental number line when estimating values, ordering numbers, judging proportions, or performing calculations (Dehaene, 2001; Schneider, Grabner, & Paetsch, 2009).

The body of evidence supporting the mental number line has been primarily provided by cognitive and developmental researchers. When comparing magnitudes, participants are quicker to discriminate between numbers that are farther apart, dubbed the distance effect (Schneider et al., 2009). Latency in comparing numbers of similar magnitude supports the holistic view of processing numerals, that numbers are judged as whole magnitudes (Dehaene, Dupoux, & Mehler, 1990; Moyer & Landauer, 1967). This contrasts with the symbolic view which supposes that numbers are processed by each digit place. To elaborate, the symbolic view posits that the ones-digit should have no effect on reaction time when comparing numerals with different tens-digits. In an example of comparing "12" and "23," the symbolic view asserts only the tens-digit is necessary and processed to judge magnitudes. Evidence, however, supports that responses are not uniform across decades and that numerals are considered as a holistic unit. In other words, the symbolic view supposes that response times when comparing "12" versus "23" and "19" versus "23" should be similar because the tens-digits are the same in both sets. Instead, evidence shows that respondents would be faster in comparing the first set due to the greater distance between the numbers. Respondents are thought to reference the two numbers against their mental number line and come to a judgment more quickly when the numbers are farther apart.

In young children, their mental number line more closely resembles a logarithmic, rather than linear, relationship (Berteletti, Lucangeli, Piazza, Dehaene, & Zorzi, 2010; Siegler & Opfer, 2003; Siegler, Thompson, & Opfer, 2009). Young children's responses to placing numerals on a number line demonstrate a linear relationship for small quantities with which children are highly familiar, such as numbers 0-10. Placing numbers outside of this familiar range results in a logarithmic pattern with greater numbers (such as 29, 42, and 56, for example) being placed relatively close together on the right end of the number line. It is not until middle elementary when students transition to a more accurate wholly-linear mental number line (Siegler et al., 2009). This phenomenon would affect children's behaviors in responding to a number line task.

Another behavioral indicator of the mental number line is the spatial–numerical association of response codes (SNARC) effect. The SNARC effect is observed through decreased latency in physical responses that are aligned to the orientation of the mental number line (Schneider et al., 2009). The mental number line is oriented with smaller quantities on the left and growing to larger quantities on the right. In alignment with this orientation, participants are quicker to respond when presented with lower quantities on the left and higher quantities on the right (Dehaene, Bossini, & Giraux, 1993; Dehaene et al., 1990; Wood, Willmes, Nuerk, & Fischer, 2008).

The mental number line also exhibits a strong relationship with overall mathematics competence (Barth & Paladino, 2011; Boyer, Levine, & Huttenlocher, 2008; Friso-van den Bos et al., 2015; Siegler, 2016). Siegler and Booth (2004), for example, found individual differences in accuracy of number line estimations correlated strongly (r = -.60 to -.76) with math achievement test scores on the Stanford Achievement Test (SAT– 9) for first and second graders. Booth and Siegler (2006) extended this work with kindergarten through fourth grade students, again finding a strong correlation between number line estimation and comprehensive math achievement test scores (ranging from r = .54 to .84). Performance on the NLT is associated with performance on magnitude comparison tasks, understanding of fractions, and overall mathematics achievement, even after controlling for compounding variables like working memory or fact fluency (Booth & Siegler, 2006; Geary, 2011; Hansen, 2015; Hansen, Jordan, & Rodrigues, 2017; Jordan et al., 2013; Schneider et al., 2018; Siegler et al., 2011). Students improve on the NLT across broad age and mathematical proficiency ranges, suggesting its utility across time (Siegler & Booth, 2004; Siegler & Opfer, 2003).

Performance on the NLT being associated with general mathematics performance suggests the potential utility of the NLT as a screening measure. Schneider et al.'s (2018) meta-analysis examined the NLT's ability to predict general mathematical competence across 41 studies with 263 effect sizes and 10,576 participants. Schneider et al. (2018) found that the average correlation (from a sample of 263 studies) between the NLT and general mathematical competence measures was r = .443 across studies. In the same meta-analysis, magnitude comparison tasks – a common task in current screening batteries – had an average correlation of r = .274 with general mathematical competence. Similar results were found (r = .438 and .278, respectively) when examining only early elementary students (aged 6-9), as well as across other age ranges, task stimuli, and methodological variations. Most importantly, a correlation of r = .443 suggests that 19.6% (r^2) of the variance in students' general mathematical performance is explained by performance on the NLT. Including the NLT within an established screening battery may aid the decisions schools make in predicting risk and serving students.

Despite the potential of the NLT as an educational tool, the number line has been primarily assessed by cognitive and developmental researchers. Prior studies on the number line have typically assessed this construct via the number line estimation task (NLT; Berteletti, Lucangeli, Piazza, Dehaene, & Zorzi, 2010; Geary, Hoard, Nugent, & Byrd-Craven, 2008; Laski & Siegler, 2007; R. Siegler & Booth, 2004; R. S. Siegler & Opfer, 2003). During typical procedures for the NLT, students are presented with a blank number line and asked to place target numerals along the line. The value of the endpoints (e.g. 0 and 20 or 0 and 100), the presence of one or both endpoints, and anchor numbers (e.g. 10, 25, 50) vary across manifestations of the task (Schneider et al., 2018). Student performance is often calculated one of three ways: (1) by summing the absolute distance of the student's responses from the correct placements, (2) by percentage of correct trials where a correct response is within a range around the correct placement, or (3) by calculating the correlation of student responses to correct placements (Schneider et al., 2018).

How students approach the NLT may also provide useful information. Descriptive analyses of the NLT suggest successful performance requires the integration of various mathematics domains. Participants must be able to, at the very least, identify the target numeral, understand the scope of the number line, and accurately estimate the numeral's place along the line. Respondents also attack stimuli differently, relying upon anchors, rounding, fractions, counting, proportional reasoning, or other strategies to produce accurate responses (Ashcraft & Moore, 2012; Peeters, Degrande, Ebersbach, Verschaffel, & Luwel, 2016; Siegler, 2016; Siegler & Opfer, 2003). Whereas one student may partition the line into salient anchors (25, 50, and 75), another may round the target stimuli to a more familiar number (12 to 10). A third may transform the stimuli's placement into a more familiar proportion (71/100 to 3/4), while a fourth may find a familiar unit and iterate along the line to estimate the target. These examples underscore how the NLT requires the simultaneous application of various mathematical skills (Siegler, 2016).

In light of its promise for educational purposes, emerging research has explored the NLT as a screening tool (Clarke, Strand Cary, Shanley, & Sutherland, 2018). Clarke et al. (Clarke et al., 2018) administered an early numeracy screener (Assessing Student Proficiency of Early Number Sense; ASPENS) and two versions of the NLT (0-20 and 0-

100) to exiting students in kindergarten (*n* = 46) and first grade (*n* = 60) as part of a five-week summer school program. They found that the NLT explained 13% additional variance above and beyond the typical screening battery for first grade students. The NLT explained 7% additional variance for kindergarten students, although this was not statistically significant. Due to the sample being drawn from a summer school program serving lower-performing students, the general population of kindergarten and first grade students was not represented. Additionally, the limited time between pre- and post-assessment limits the ability of these results to generalize to fall screening processes in schools.

Sutherland and colleagues (2020) expanded on the prior study in drawing upon a broader sample (*n* = 117) of kindergarten students in control classrooms from a larger study, representing a general population. Additionally, measures were administered in the fall and spring, approximating typical screening and outcome processes. Due to time constraints, administration of the number line measure ended after five minutes. The number line assessment (0-100) performed similarly to the typical mathematics screener (ASPENS; *r* = .60 and *r* = .62, respectively). Independently, the ASPENS explained 49% of students' spring mathematics performance while the NLT explained 35%. Additionally, the ASPENS exhibited an Area Under the Curve (AUC) value of .94 compared to the .80 of the NLT. When considered in combination, however, the NLT uniquely explained 7% of the variance in spring mathematics performance above and beyond the ASPENS. Across studies, the NLT results indicate potential value as a supplement to, but not necessarily a wholesale replacement of, established mathematics screening batteries.

In both of the preceding studies, the NLT was adapted from Laski et al.'s (2013) version of 26 items. No screening study to date has investigated whether a form of less items, allowing for greater efficiency, could provide comparable predictive value. The goal for practice is to increase or preserve the diagnostic accuracy of a screener while optimizing its efficiency. Some evidence suggests that reducing the number of items on selected measures may be an alternative. For example, Purpura and colleagues (2015) found comparable information was garnered between their original 143-item screening battery and the shortened form of 24 items. Similarly, Rodrigues and colleagues (2019) reduced their screening net by 38-39 items and up to an estimated 29 minutes of administration time while increasing predictive power. By removing items that do not contribute to the overall prediction power of the measures, comparable information can be captured in less time.

The mental number line is a central conceptual structure that children form as they grow acquainted with quantities and that develops as children do to eventually accommodate more advanced numbers (Siegler et al., 2011). As a measure that assesses a central construct across years of mathematics, the NLT provides numerous conceptual arguments for exploration as a screener.

State of the Problem: The Current Study

Despite the importance of mathematics for academic and occupational success, students in the United States are not mastering critical content. Universal screening presents an opportunity for schools to wisely leverage resources and identify students most in need of additional support. Early mathematics research is lacking consensus on best practices in screening. The mental number line offers promise as a screener to accurately and efficiently identify students at risk. As part of a larger study, 226 students were administered the NLT in the fall along with an established early numeracy screener, a short outcome measure, and a comprehensive outcome measure. Analyses explored the predictive properties of a short form four-item NLT. The NLT was compared to an established screener for the extent to which it added value in predicting performance on the spring mathematics outcome measures. Lastly, the items within the NLT were explored for differences in utility.

Research Questions

1. To what extent does the Number Line Task (NLT) predict math performance in an educational context?
    A. What are the associations among the NLT and other measures of early numeracy?
    B. To what extent does the NLT add value above and beyond a typical mathematics screener (the ASPENS)?
    C. What are the classification accuracy statistics of the NLT compared to an established math screener (ASPENS)?
2. Within the NLT, which items explain the most variance? Do items differ in their utility for decision-making?

II. METHOD

This study analyzed data from a larger randomized controlled trial examining the efficacy of the federally-funded ROOTS kindergarten mathematics intervention program (Clarke, Doabler, Fien, Baker, & Smolkowski, 2012). ROOTS is a 50-lesson intervention program that focuses on improving student understanding of whole number concepts and associated skills.

Math achievement data were collected at the individual level for students. Random assignment and instructional delivery took place at the classroom level. Blocking on school and teacher experience with the core curriculum (one year or none), classrooms were randomly assigned to treatment and control conditions. Assessments were administered in the fall and spring of kindergarten.

Participants

Participants were drawn from the first cohort of the parent study (Clarke et al., 2012). Participants were 226 kindergarten students from 14 classrooms during the 2012-2013 schoolyear. The classrooms were nested within 7 schools within 3 districts. From the 785 students of the parent study, the final sample for analysis (*n* = 226) removed cases with partial or complete missing data (*n* = 559) for all measures (fall NLT, fall ASPENS, fall NSB, spring ASPENS, spring NSB, and spring SESAT).

Welch independent two-sample *t*-tests were conducted to determine if student fall mathematics scores differed for included students as compared to excluded students with available fall data. There were no significant differences on any of the fall mathematics measures; NLT *t*(42.91) = 1.61, *p* = .21. ASPENS, *t*(45.38) = 0.22, *p* = 0.64, and NSB, *t*(45.36) = .01, *p* = 0.91.

Similarly, Welch independent two-sample *t*-tests were conducted to determine if student's fall mathematics scores differed for students assigned to intervention as compared to students assigned to the control condition. There were no significant differences on any of the fall mathematics measures; NLT *t*(61.95) = .55, *p* = .46. ASPENS, *t*(64.39) = 0.05, *p* = 0.82, and NSB, *t*(84.59) = .25, *p* = 0.62.

Participating school districts were all in suburban and rural areas of western Oregon. Schools targeted for recruitment across the three districts were primarily those that received Title 1 funding. Of the 226 students in the sample: 129 (57.1%) were female; 150 (66.4%) were 5-years-old, 76 (33.6%) were 6-years-old, 196 students (86.7%) were White, 7 students (3.1%) were American Indian or Alaskan Native, 7 students (3.1%) were Black or African American, 30 students (13.3%) identified as Hispanic and/or Latino, and five or fewer students identified as Asian, Native Hawaiian/Pacific Islander, or more than one race; 15 students (6.6%) were English learners; and 15 students (6.6%) received special education services.

Procedures

Students were individually administered all measures by trained staff with extensive experience in collecting data for educational research. Interrater reliability of all administrators was at least .90 before collecting data with students. Administrators attended follow-up trainings prior to data collection sessions to prevent drift from standardization.

Student assessment protocols were processed using Teleform, a form processing application. Tests of Teleform scoring procedures of assessment protocols from previous research projects reveal high reliability values (i.e., .99) relative to assessor-scored protocols (.95).

Measures

Four of the parent study's mathematics measures were chosen to investigate the research questions in this sub-study: the 0-100 NLT, ASPENS, NSB, and SESAT.

Number Line Estimation Task (NLT)

Administrators folded a paper in half lengthwise and handed the paper and a red pencil to the student. Administrators said, "This is a number line. If this is 0 and this is 100 (administrator points to each endpoint while talking), where would 34 be? Use the pencil, and mark on the number line where 34 would be." Each page displayed two number lines but was folded so that the student would only see one number line at a time. The administrator then displayed a new number line and prompted the student for next item, asking, "where would [x] be?" Items were 34, 12, 89, and 57.

Responses were scored as the absolute distance of the students' responses from the correct responses. The first mark a student places on the number line was used for scoring. A transparency was laid over the student's response form and the administrator counted the spaces between the student's response and the target number. For example, a student was told to locate 34 and marked the number line where 42 resides. This student received a score of 8. Summed scores closest to zero indicate better performance. The four stimuli were chosen semi-randomly, sampling across the range of 0 to 100.

Assessing Student Proficiency of Number Sense (ASPENS; Clarke, Gersten, Dimino, & Rolfhus, 2011)

The ASPENS is a series of three one-minute curriculum-based measures of numeral identification, comparing quantities, and strategic counting. The ASPENS is utilized in the present study as a proxy for a typical mathematics screening battery due to including magnitude comparison and strategic counting tasks. Test-retest reliabilities of kindergarten ASPENS measures are in the moderate to high range (.74 to .85). Predictive validity from fall to spring scores on the TerraNova 3 is reported as ranging from .45 to .52.

Number Sense Brief (NSB; Jordan, Glutting, & Ramineni, 2008)

The NSB is an individually administered measure with 33 items drawing upon varied early numeracy skills, such as counting knowledge and principles, number recognition, number comparisons, nonverbal calculation, story problems, and number combinations. The NSB has a coefficient alpha of .84. The NSB serves as a short outcome measure for determining general student mathematics performance.

***The Stanford Early School Achievement Test – Tenth Edition* (SESAT;** Harcourt Educational Measurement, 2002)

The Stanford Early School Achievement Test – Tenth Edition (SESAT) is a group-administered standardized, norm-referenced achievement test with two multiple-choice mathematics subtests, Problem Solving and Procedures. The SESAT has adequate validity (*r* = .67) and reliability (*r* = .93). The SESAT serves as a longer, comprehensive outcome measure for determining general student mathematics performance.

Analyses

Prior to analyses for the study's research questions, univariate descriptive statistics for each measure at each timepoint were calculated and assumptions of fitness for linear regression (linearity, independence of errors, multivariate normality, and homoscedasticity) were tested (Pedhazur & Kerlinger, 1982).

To address research question 1A, Pearson's *r* bivariate correlations were estimated among the NLT, established early numeracy screener (ASPENS), and outcome measures (NSB and SESAT).

To address research question 1B, six linear regression models were conducted. Table 1 displays the conducted models. In the first model, the spring NSB scores were regressed on the fall NLT scores. In the second model, spring NSB scores were regressed on the fall ASPENS scores. In the third, the spring NSB scores were regressed on both predictors, fall NLT and ASPENS scores. This procedure was repeated for the remaining three models by regressing the second outcome measure, the spring SESAT scores, on the same set of predictors. Including the intervention condition in the combined models only slightly increased predictiveness (by $R^2 = .01$ and .02 for the spring NSB and spring SESAT, respectively). For the sake of parsimony, intervention condition was excluded from the models. The R^2 value given by each model estimates the variance explained by the predictors in the outcome measures. Semi-partial correlations were estimated for the final model for each outcome measure. Semi-partial correlations parse out shared variance to better understand what each independent variable uniquely contributes to the model.

Table 1

Linear Regression Models Conducted for Research Question 1

Model Number	Predictor(s)	Outcome
1A	Fall NLT	Spring NSB
1B	Fall ASPENS	Spring NSB
1C	Fall NLT + Fall ASPENS	Spring NSB
2A	Fall NLT	Spring SESAT
2B	Fall ASPENS	Spring SESAT
2C	Fall NLT + Fall ASPENS	Spring SESAT

To address research question 1C, receiver operating characteristic (ROC) analyses assessed the diagnostic accuracy of the NLT and the ASPENS. ROC analyses evaluate a measure's classification performance to a dichotomous outcome variable of "risk." A cut score of 20 on the NSB was used to qualify "risk" with those scoring above 20 deemed "not at risk." A cut score of 20 aligns with previous research supporting its utility for diagnostic accuracy in the spring of kindergarten (Jordan, Glutting, Ramineni, & Watkins, 2010). Additionally, a score of 20 corresponds to the 23rd percentile in this sample, which holds clinical significance and approximates a threshold schools may use to assign intervention. For this reason as well, performance below the 25th percentile on the SESAT in this sample was deemed as "at risk."

When evaluating a ROC curve, the area under the curve (AUC) estimates how well a measure accurately classifies subjects. Values close to 1 suggest a measure is highly sensitive and specific (or accurately parses out individuals who are truly "at risk" or truly "not at risk"). Values close to .5, in contrast, denote the measure performs little better than chance. Confidence intervals and statistically significant differences for the AUC values were computed using 2,000 stratified bootstrap replicates.

To address the second research question, analyses mirrored the procedure of the first research question. Pearson's *r* bivariate correlations were estimated among the individual NLT items and the ASPENS, NSB and SESAT at all available timepoints. Next, each outcome measure (the spring NSB or spring SESAT) was regressed on the individual fall NLT item scores and the fall ASPENS scores. As with research question 1B, the R^2 value and semi-partial correlations were collected. Lastly, AUC values were conducted for the NLT Items to examine classification accuracy.

It was hypothesized that Item 4 (with a stimulus of 57) would be associated with greater overall mathematical competence in kindergarten. Basis from this hypothesis drew from Rodrigues, Jordan and Hansen (2019) who found that "simpler" items such as the midpoint (1/2) on a 0-1 fraction number line were the most predictive items. Similarly, it was hypothesized that a stimulus of 57 would require students to demonstrate foundational skills in a) identifying the two-digit numeral correctly and b) dissecting the line approximately in half. Exhibiting these developing mathematical competencies may be associated with greater overall mathematical competence in kindergarten.

Type I error rate for all analyses was set at 5% (.05) as is standard in educational sciences. All analyses were conducted in R (R Development Core Team, 2011), with the following packages: cowplot (Wilke, 2019); ggplot2 (Wickham, 2016); ggResidpanel (Goode & Rey, 2019); ggROC (Wu, 2013); haven (Wickham & Miller, 2019); here (Müller, 2017); Hmisc (Harrell Jr, 2020); lmSupport (Curtin, 2018); pROC (Robin et al., 2011); rio (Chan, Chan, Leeper, & Becker, 2018); and tidyverse (Wickham et al., 2019).

III. RESULTS

Univariate descriptive statistics are displayed in Table 2. Students gained, on average, about 44 points (77.5%) from fall to spring of kindergarten on the ASPENS measure. Students also gained, on average, on the NSB measure by about 5.6 points (31.9%).

Table 2

Descriptive Statistics of Early Numeracy Measures

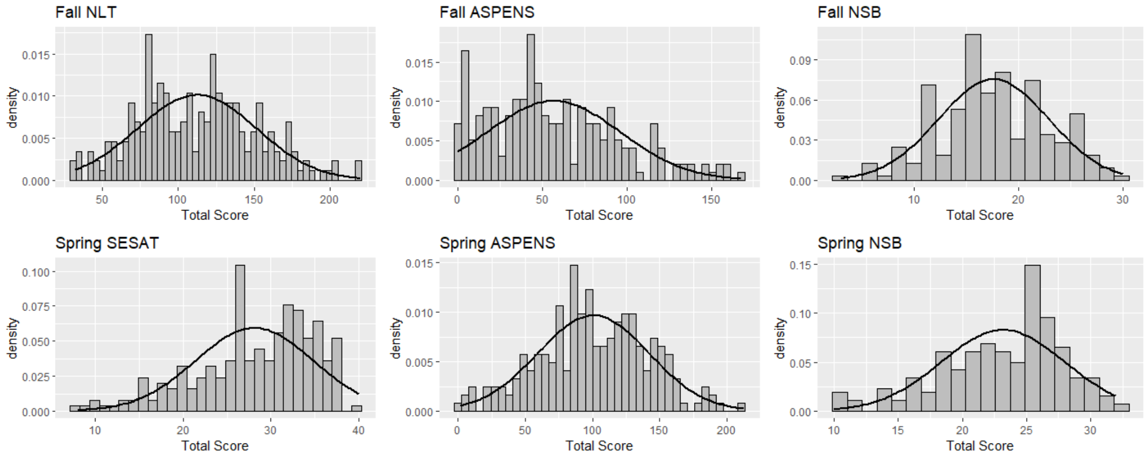
Measure	Mean	SD	Median	Skewness	Kurtosis
Fall NLT	112.27	39.40	112.50	0.22	-0.45
Fall ASPENS	56.73	39.66	48.85	0.64	-0.21
Fall NSB	17.58	5.29	17.00	-0.02	-0.46
Spring ASPENS	100.65	41.24	98.80	-0.07	-0.28
Spring NSB	23.17	4.81	24.00	-0.61	-0.17
Spring SESAT	28.11	6.72	29.00	-0.77	0.00

Assumptions of fitness for linear regression were tested. First, the variables were examined for normality. Distributions of the study measures are displayed in Figure 1. The fall NLT scores approximate a normal distribution (Shapiro-Wilk normality test *p* = .07). The fall NLT scores have a slight positive skew (0.22) and less kurtosis than expected (-0.45). However, graphical representation suggests that the distribution of fall NLT scores may be bimodal. The fall ASPENS scores fail the Shapiro-Wilk normality test (*p* < .001). The fall ASPENS scores have moderate positive skew with approximately normal kurtosis (-0.21). The spring ASPENS scores, however, do approximate a normal distribution (Shapiro-Wilk normality test *p* = .63), with minimal skew (-0.07) and expected kurtosis (-0.17).

The fall NSB scores approximate a normal distribution (Shapiro-Wilk normality test *p* = .11), with minimal skew (-0.02) and less kurtosis than expected (-0.46). In contrast, the spring NSB scores fail the Shapiro-Wilk normality test (*p* < .001). The spring NSB scores demonstrate moderate negative skew (-0.61) and approximately normal kurtosis (-0.17). The spring SESAT scores fail the Shapiro-Wilk normality test (*p* < .001). The spring SESAT scores demonstrate moderate negative skew (-0.77) and expected kurtosis (0.00).

Figure 1

Distributions of the Early Numeracy Measures



Note. Axes' scales vary by measure.

Normal quantile plots were examined to determine multivariate normality. Quantile plots are displayed in Figures 2 and 3. The linear trend displayed by the theoretical quantities plotted against the sample quantities predicting to the NSB suggests multivariate normality. The tails of the model predicting to the SESAT (Figure 3) deviate from a linear trend. These deviations suggest the multivariate distribution predicting to the SESAT is negatively skewed.

Figure 2

Normal Quantile Plot of Spring NSB Scores Regressed on Fall NLT and Fall ASPENS Scores

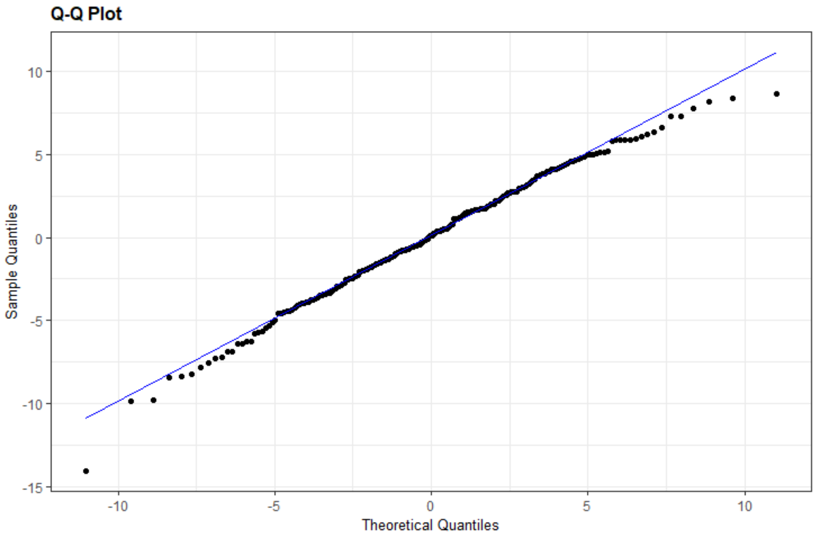
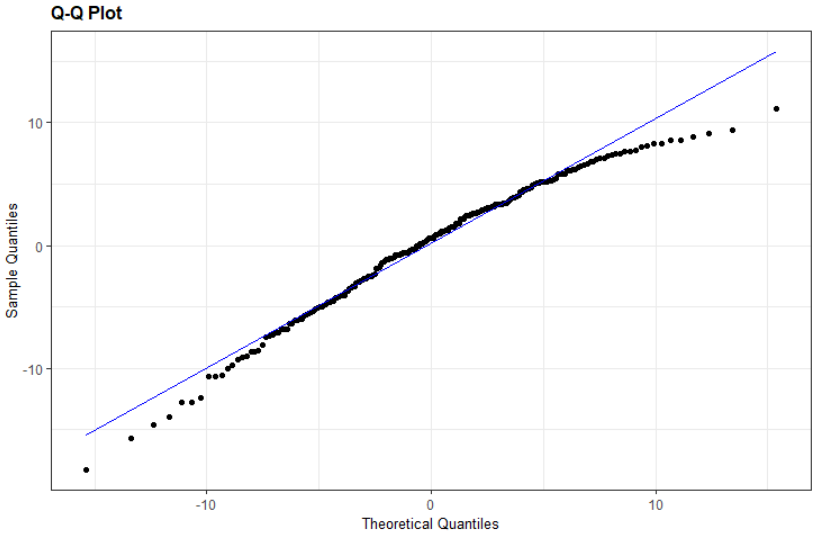


Figure 3

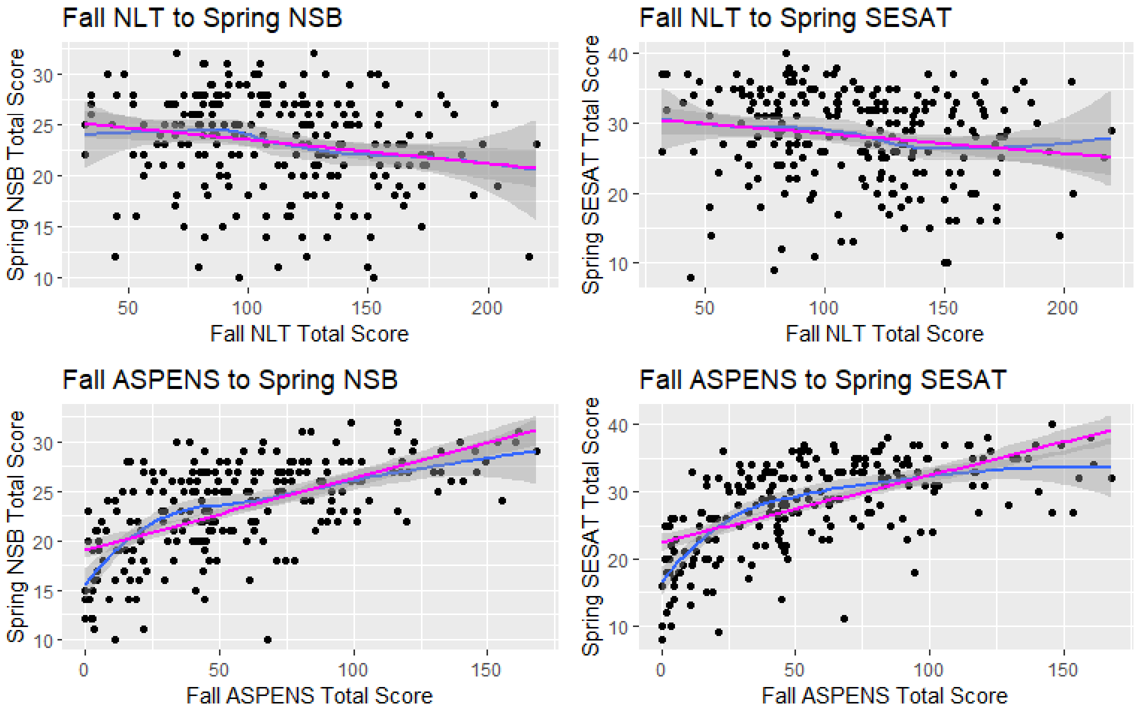
Normal Quantile Plot of Spring SESAT Scores Regressed on Fall NLT and Fall ASPENS Scores



Next, linearity of the predictor models was examined. The spring outcome measures (NSB and SESAT) were each regressed on the fall predictor measures (NLT and ASPENS). Linear regressions are displayed in Figure 4. In all models, a linear trend appears to best explain the relationship between the predictors and the outcomes. The assumption of linearity is tenable.

Figure 4

Simple Linear Regressions of the Outcomes Regressed on the Predictor Measures



The assumption of independence of errors was examined next. Residuals of the outcomes regressed on the predictors are plotted in Figures 5 and 6. Residuals for the NLT appear randomly distributed, suggesting the absence of a relationship between the errors and the outcome variables. Residuals for the ASPENS predicting to each outcome appear to be somewhat overestimated at the extreme values and underestimated at the central values.

Figure 5

Residual Plots of the Spring NSB Scores Regressed on the Fall NLT scores and Fall ASPENS scores

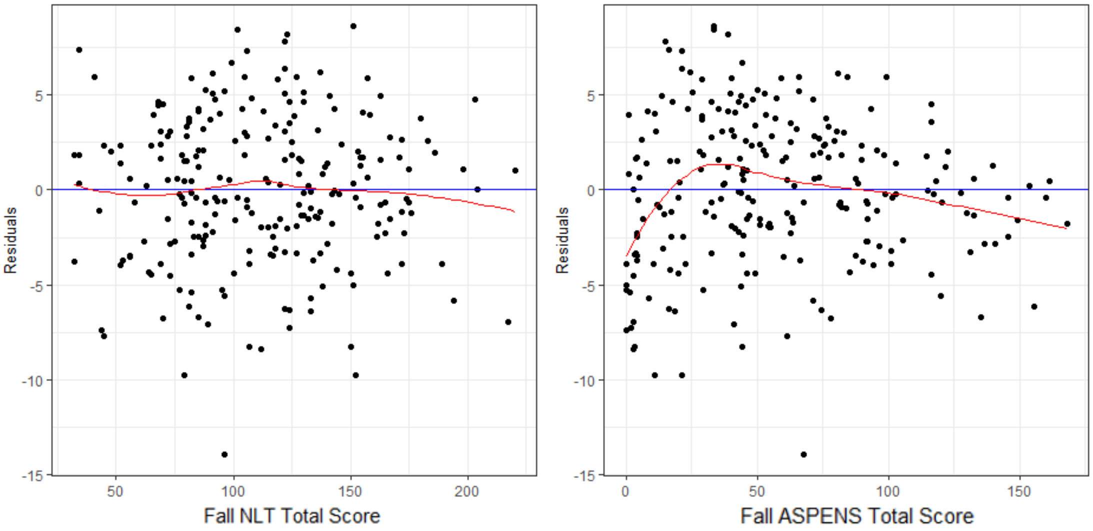
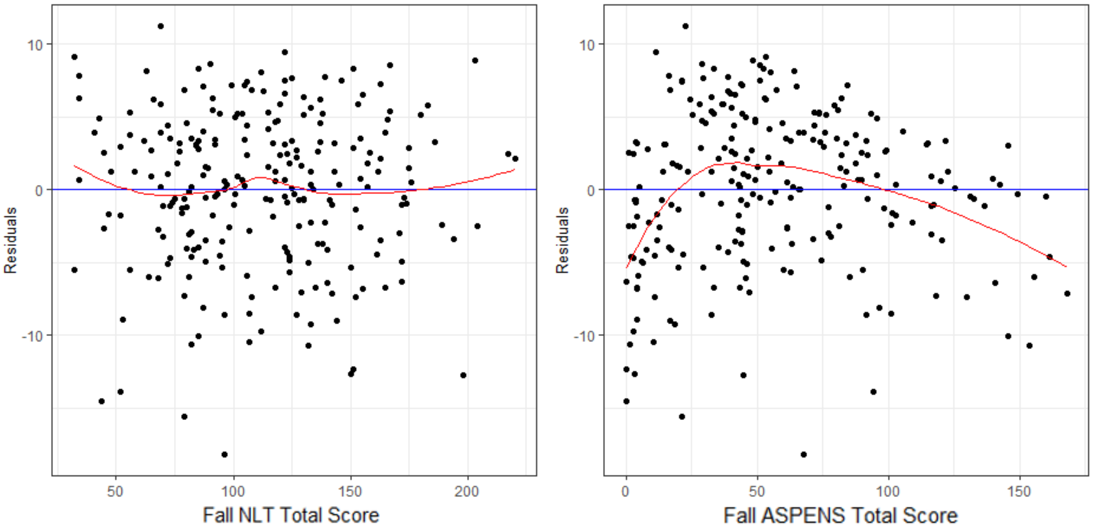


Figure 6

Residual Plots of the Spring SESAT Scores Regressed on the Fall NLT scores and Fall ASPENS Scores



Lastly, the assumption of homoscedasticity was examined. Further examination of the residual plots shows that, for the fall NLT as a predictor, errors appear homogenously distributed across the values of the x-axes. The assumption of homoscedasticity is tenable for the spring NSB and spring SESAT regressed on the fall NLT. The residuals of the fall ASPENS as a predictor do not appear homogenously distributed across the values of the x-axes. The variance of the residuals tend to decrease going across the x-axes. The assumption of homoscedasticity for the fall ASPENS is not tenable.

Research Question 1

Research Question 1A: Association Among Early Numeracy Measures

Correlations among all measures at all available timepoints were conducted. Descriptors of the strength of correlations are based on Cohen (1992) who defines small, medium and large correlations as *r* = |.20|, |.30|, and |.50|, respectively. Correlations for all measures are reported in Table 3. Correlations for the study measures only (Fall NLT, Fall ASPENS, Spring NSB, and Spring SESAT) are displayed graphically in Figure 7. Associations with the NLT are expected to be negative as larger scores indicate greater error (response distance from the target numbers).

Except for the relations of the fall NLT with the spring ASPENS and fall NLT with the spring SESAT, all correlations are significant (*p* < .01). The relationship between the NLT and the ASPENS in the fall is small (*r* = -.26) and weak in the spring (*r* = -.13, *p* = .058). The relationship between the NLT and the NSB is small in the fall (*r* = -.24). In the spring, the NLT's relationships with the NSB (*r* = -.19) and the SESAT (*r* = -.17, *p* < .05) are weak. The ASPENS in the fall is strongly correlated with the NSB in the fall (*r* = .68). The ASPENS remains strongly associated with the NSB in the spring (*r* = .67) and with the other outcome measure administered in the spring, the SESAT (*r* = .64). The outcome measures demonstrate strong relationships among each other at all timepoints (*r* = .66 to .73), echoing prior evidence of validity (Harcourt Educational Measurement, 2002; Jordan et al., 2008).

Table 3

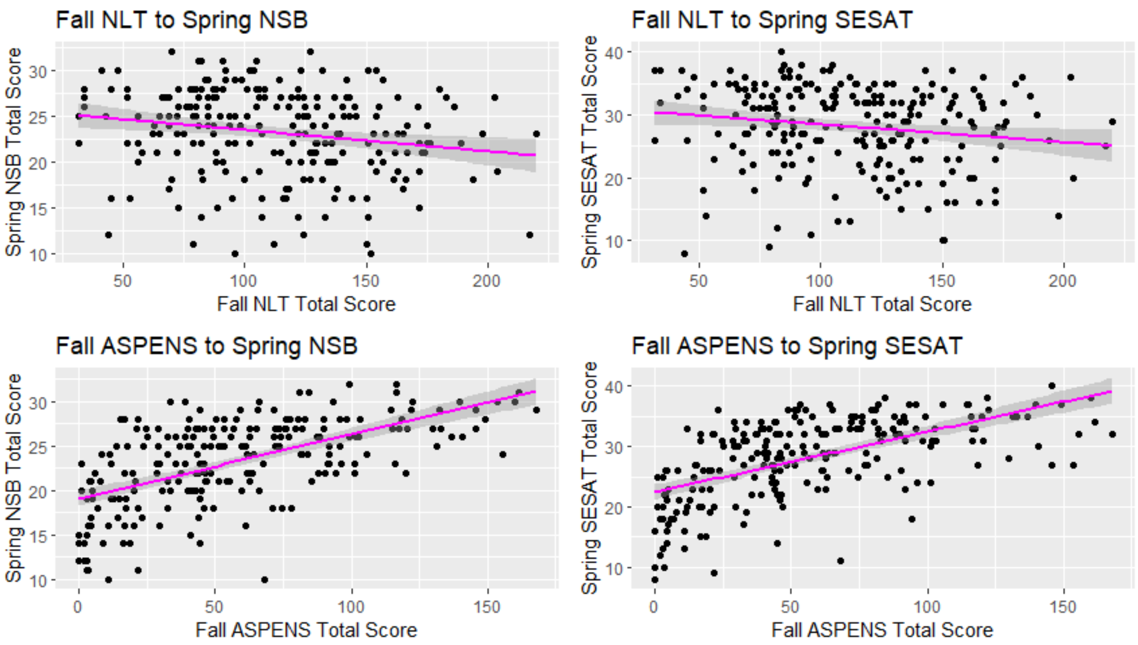
Correlations Among All Measures Administered Fall 2012 and Spring 2013

	1	2	3	4	5	6
1. Fall NLT	-	-.26**	-.24**	-.13	-.19**	-.17*
2. Fall ASPENS		-	.68**	.70**	.59**	.59**
3. Fall NSB			-	.57**	.72**	.66**
4. Spring ASPENS				-	.67**	.64**
5. Spring NSB					-	.73**
6. Spring SESAT						-

**p* < .05. ***p* < .01.

Figure 7

Correlations Among Study Measures Administered Fall 2012 and Spring 2013



Research Question 1B: Explained Variance

Results of the outcome measures regressed on the predictor measures are reported and summarized. In the first model, the spring NSB scores were regressed on the fall

NLT scores. In the second model, spring NSB scores were regressed on the fall ASPENS scores. In the third, the spring NSB scores were regressed on both predictors, fall NLT and ASPENS scores. This procedure was repeated for the remaining three models by regressing the second outcome measure, the spring SESAT scores, on the same set of predictors. Regression results are reported in Tables 4 and 5.

Fall performance on the NLT explained 3% of the variance in scores on the spring NSB scores, as well as for spring SESAT scores. Students' performance on the fall ASPENS explained 35% of the variance in scores on the spring NSB, and likewise for the spring SESAT scores. Models that included both predictors (fall NLT and fall ASPENS scores) did not show an increase in explained variance in the outcomes over the models that included only the ASPENS. In addition, the fall NLT scores were no longer a statistically significant predictor in the combined models (p = .48 predicting to the spring NSB, p = .83 predicting to the spring SESAT).

In the combined model for predicting the spring NSB (Model 3), for every 1-point increase in the NLT, there is no expected increase in spring NSB score (p = .48). For every 1-point increase on the fall ASPENS, there is an expected .07 increase in score on the spring NSB (p < .001). This model accounts for approximately 35% of the variance in scores on the spring NSB, F(2, 223) = 59.52, p < .001.

In the combined model for predicting the spring SESAT (Model 6), for every 1-point increase in the NLT, there is no expected increase in spring SESAT score (p = .83). For every 1-point increase on the fall ASPENS, there is an expected .10 increase in score on the spring SESAT (p < .001). This model accounts for approximately 35% of the variance in scores on the spring SESAT, F(2, 223) = 59.54, p < .001.

Semi-partial correlations explain the extent to which each predictor adds unique variance in explaining the outcome, net of the shared variance among predictors. Semi-partial correlations are reported in Table 6. In the combined models (Models 3 and 6), the NLT adds negligible explained variance (R^2 <.01) in the outcome measures beyond the ASPENS.

Table 6

Unique Variance Explained in the Outcome Measures (Semi-partial Correlations)

	Spring NSB	Spring SESAT
Fall NLT	<.01	<.01
Fall ASPENS	.31*	.32*

*p < .01.

Research Question 1C: Classification Accuracy

ROC analyses explored the predictors' abilities to correctly classify students at risk. The ROC curves predicting to the spring NSB and spring SESAT are displayed in Figures 8 and 9. AUC values are reported in Table 7.

The AUC of the fall NLT to the spring NSB was .59 (95% CI from .49 to .68) and to the spring SESAT was .58 (95% CI from .49 to .66). The AUC of the fall ASPENS to the spring NSB was .86 (95% CI from .80 to .92) and to the spring SESAT was .83 (95% CI from .76 to .89). For both outcome measures, the fall ASPENS greatly outperformed the NLT in accurately classifying students. These differences are statistically significant for both predicting to both measures (p < .001).

Table 4

Regression Results Predicting Spring NSB Performance (N = 226)

	Model 1				Model 2				Model 3			
Parameter	*b*	*SE*	*T*	*p*	*b*	*SE*	*t*	*p*	*b*	*SE*	*t*	*p*
Intercept, b_1	25.80	0.95	27.08	<.001	19.12	0.45	42.18	<.001	19.73	0.98	20.11	<.001
Fall NLT, b_2	-0.02	0.01	-2.92	<.01					0.00	0.01	-0.70	.48
Fall ASPENS, b_3					0.07	0.01	10.90	<.001	0.07	0.01	10.32	<.001

Note. Model 1 R^2 = .03, F = 7.95, p = .01. Model 2 R^2 = .35, F = 111.80, p < .001. Model 3 R^2 = .35, F = 59.52, p < .001.

Table 5

Regression Results Predicting Spring SESAT Performance (N = 226)

	Model 4				Model 5				Model 6			
Parameter	*b*	*SE*	*T*	*p*	*b*	*SE*	*t*	*p*	*b*	*SE*	*t*	*p*
Intercept, b_1	31.29	1.34	23.42	<.001	22.44	0.63	35.52	<.001	22.70	1.37	16.58	<.001
Fall NLT, b_2	-0.03	0.01	-2.52	.01					0.00	0.01	-0.21	.83
Fall ASPENS, b_3					0.10	0.01	10.93	<.001	0.10	0.01	10.47	<.001

Note. Model 1 R^2 = .03, F = 6.36, p = .01. Model 2 R^2 = .35, F = 119.60, p < .001. Model 3 R^2 = .35, F = 59.54, p < .001.

Figure 8

ROC Curve Comparing Fall NLT and Fall ASPENS to the Spring NSB

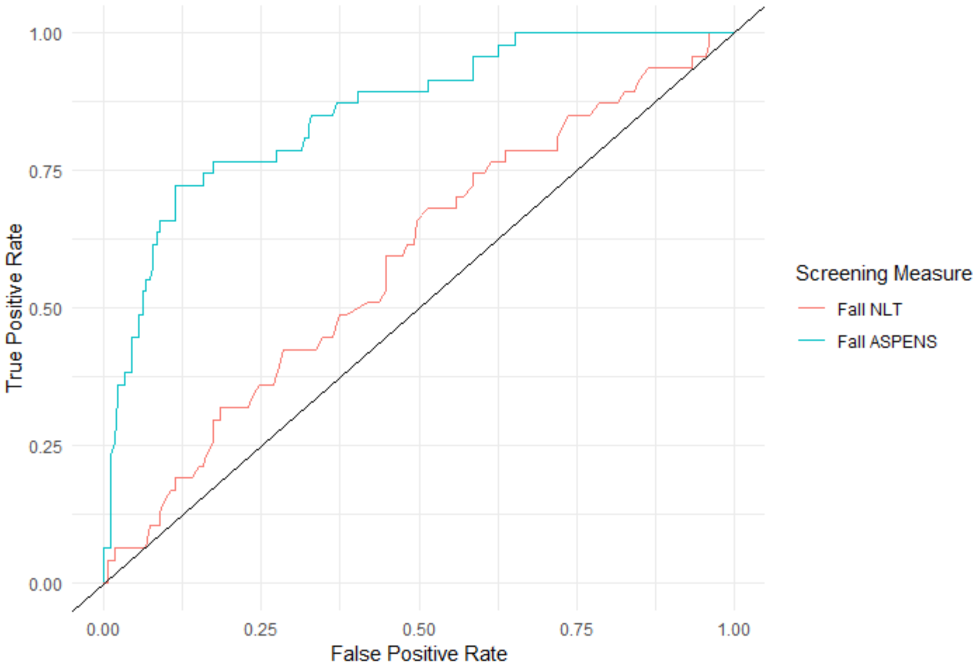


Figure 9

ROC Curve Comparing Fall NLT and Fall ASPENS to the Spring SESAT

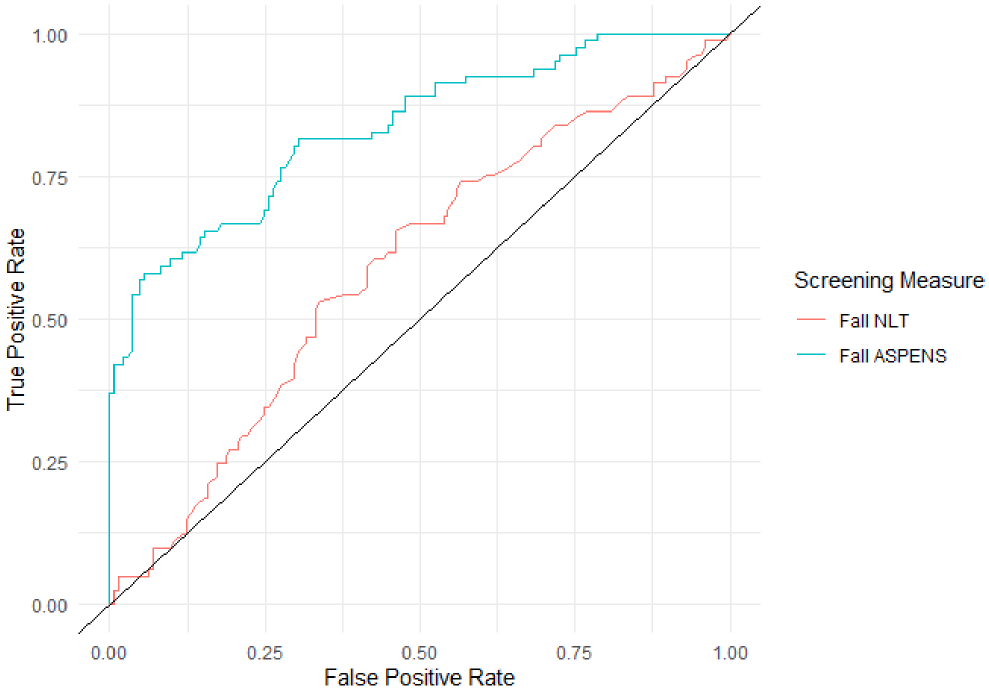


Table 7

AUC for the Fall Screening Measures to the Spring Outcome Measures

	Spring NSB (23rd Percentile)		Spring SESAT (25th Percentile)	
	AUC	CI	AUC	CI
Fall NLT	.59	.49-.68	.59	.51-.67
Fall ASPENS	.86	.80-.92	.83	.78-.89

Following the ROC analyses, sensitivity and specificity were examined. Classification accuracy statistics and cut scores are reported in Table 8. Two approaches were used. The first approach maximized both sensitivity and specificity. The cut score with the sum of sensitivity and specificity closest to 2.0 was selected for each predictor to each outcome. When risk was classified as below the 23rd percentile in this sample on the NSB, the fall NLT had a sensitivity of .68 and a specificity of .49 (cut score = 105.50) whereas the fall ASPENS had a sensitivity of .72 and a specificity of .89 (cut score = 23.35). While the measures correctly identified students "at risk" to a similar extent (4% difference in favor of the ASPENS), the ASPENS correctly identified 40% more "not at risk" students.

When risk was classified as below the 25th percentile in this sample on the SESAT, the fall NLT had a sensitivity of .53 and a specificity of .66 (cut score = 122.50) whereas the fall ASPENS had a sensitivity of .58 and a specificity of .95 (cut score = 24.25). Again, the ASPENS correctly identified slightly more "at risk" students than the NLT (5%), and substantially more "not at risk" students (29%).

Because the implications for false negatives are greater than for false positives for students, schools often prioritize sensitivity over specificity. Thus, the next approach examined cut scores and specificities where sensitivity was closest to .90 (L. S. Fuchs et al., 2007; Seethaler & Fuchs, 2010). When risk was classified as below the 23rd percentile on the NSB, the fall NLT had a specificity of .16 (sensitivity = .89, cut score = 71.00) whereas the fall ASPENS had a specificity of .60 (sensitivity = .89, cut score = 49.25). When the measures correctly identified 89% of truly "at risk" students, the ASPENS correctly identified 44% more "not at risk" students. When risk was classified as below the 25th percentile on the SESAT, the fall NLT had a specificity of .12 (sensitivity = .91, cut score = 67.00) whereas the fall ASPENS had a specificity of .48 (sensitivity = .90, cut score = 69.55). When the measures correctly identified close to 90% of truly "at risk" students, the ASPENS outperformed the NLT in correctly identifying "not at risk" students by 36%.

Table 8

Classification Accuracy and Cut Scores for Fall Screeners Maximizing Sensitivity and Specificity and with Sensitivity Closest to .90

Fall NLT			Fall ASPENS		
Cut Score	Sens	Spec	Cut Score	Sens	Spec
Spring NSB					
105.50	.68	.49	23.35	.72	.89
71.00	.89	.16	49.25	.89	.60
Spring SESAT					
122.50	.53	.66	24.25	.58	.95
67.00	.91	.12	69.55	.90	.48

Note. Sens = Sensitivity, Spec = Specificity

Research Question 2: Item-Level Analyses

Analyses were repeated with the individual NLT items to explore which stimuli best predict future achievement. Descriptive statistics of the NLT Items and the outcome measures are displayed in Table 9.

Means for the NLT Items were examined as a one-way, repeated measures analysis of variance. The independent variable was the NLT Item with four levels (Items 1 to 4) and the dependent variable was the received score. The Mauchly Sphericity test was not significant, indicating that the assumption of sphericity is tenable, $\chi^2(5) = .29$. The main effect of NLT Item on score was significant, $F(3, 900) = 6.58$, $p < .001$. Post-hoc tests using a Bonferroni correction reveal that the mean score of Item 2 is significantly greater than Items 1 and 4 ($p < .01$). Other pairwise comparisons are not significant.

Table 9

Descriptive Statistics of the NLT Items and Outcome Measures

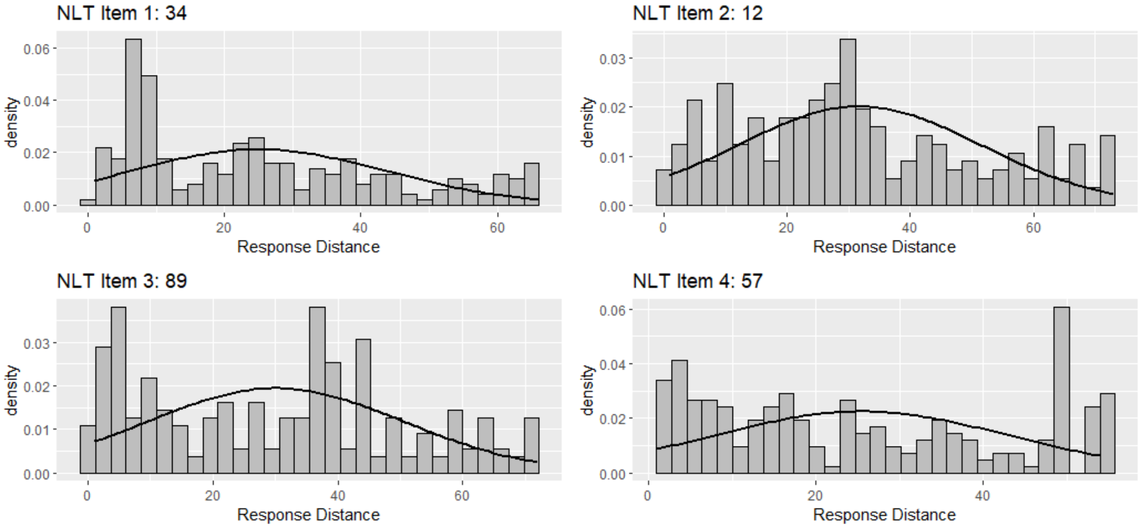
Measure	Mean	SD	Median	Skewness	Kurtosis
Fall NLT Item 1: 34	25.09^{a}	18.87	22.00	0.66	-0.75
Fall NLT Item 2: 12	31.69^{b}	19.80	29.00	0.42	-0.81
Fall NLT Item 3: 89	30.01^{ab}	20.58	32.50	0.23	-1.04
Fall NLT Item 4: 57	25.47^{a}	17.81	23.00	0.26	-1.36
Spring NSB	23.17	4.81	24.00	-0.61	-0.17
Spring SESAT	28.11	6.72	29.00	-0.77	0.00

Note. Superscripts denote significantly different group means, $p < .05$.

Next, assumptions of fitness for statistical analyses were tested. The distribution of the NLT item scores are displayed Figure 10. The distributions of the NLT Items do not approximate normality. The distributions of NLT Items 1 and 2 are moderately negatively skewed. The distribution of Item 3 appears to be bimodal. All four item distributions have significantly more kurtosis than expected.

Figure 10

Distribution of the Individual NLT Items



Next, linearity of the predictor models was examined. The spring outcome measures (NSB and SESAT) were each regressed on the NLT item scores and fall ASPENS scores. Residuals of the regressions are displayed in Figures 11 and 12. A linear trend is apparent in both models and thus the assumption of linearity is tenable in both models.

The assumption of independence of errors was tested next. Residuals for the NLT items appear randomly distributed, suggesting the absence of a relationship between the errors and the outcome variables. The assumption of homoscedasticity was considered next. In both models, the variance appears homogenously distribution across values of x. The assumptions of independence of errors and homoscedasticity are tenable.

Figure 11

Residuals of the Spring NSB Regressed on the Fall NLT Item Scores and Fall ASPENS Scores

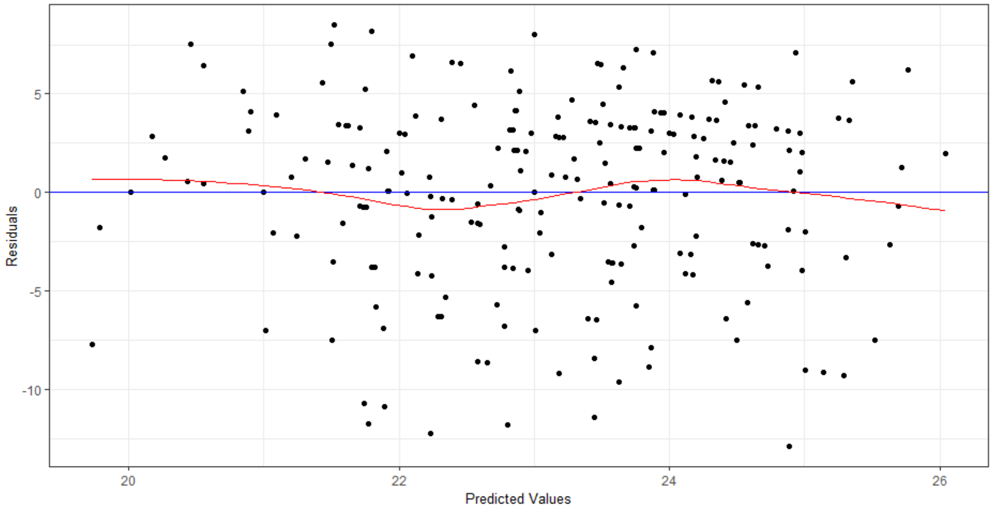
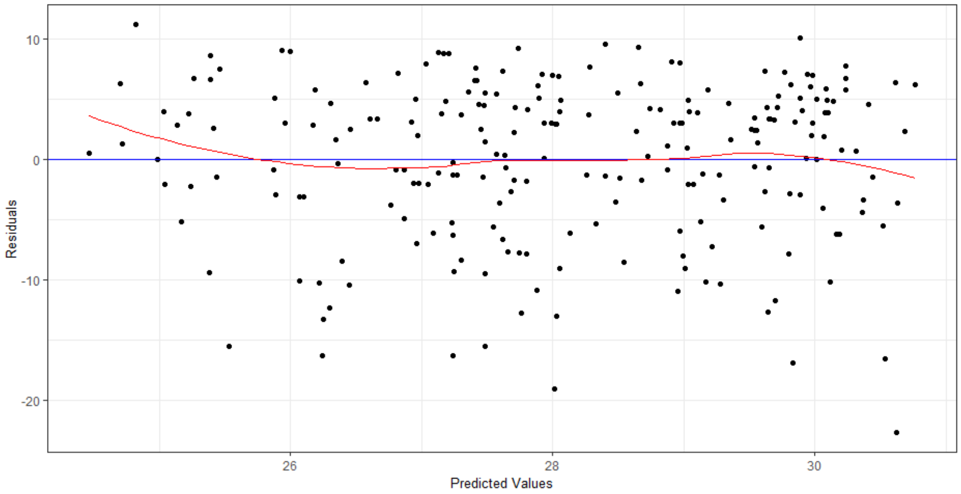


Figure 12

Residuals of the Spring SESAT Regressed on the Fall NLT Item Scores and Fall ASPENS Scores



Next, correlations among the items of the NLT and other early numeracy measures were conducted. Conducted correlations are displayed in Table 10. Items 1, 2, and 4 of the NLT are unrelated to all other items and measures (*r* < .20). In the fall, Item

Table 10

Correlations Among the NLT Items and Early Numeracy Measures

	1	2	3	4	5	6	7	8	9
1. Fall NLT Item 1: 34	-	.05	-.01	.01	-.05	-.14*	-.04	-.08	-.02
2. Fall NLT Item 2: 12		-	-.05	.03	-.14*	-.12	-.10	-.09	-.04
3. Fall NLT Item 3: 89			-	.06	-.28**	-.21**	-.16*	-.23**	-.22**
4. Fall NLT Item 4: 57				-	-.05	-.03	.06	.03	-.03
5. Fall ASPENS					-	.68**	.70**	.59**	.59**
6. Fall NSB						-	.57**	.72**	.66**
7. Spring ASPENS							-	.67**	.64**
8. Spring NSB								-	.73**
9. Spring SESAT									-

**p* < .05. ***p* < .01.

3 of the NLT has a small relationship with the ASPENS (*r* = -.28) and the NSB (*r* = -.21). This relationship continues into the spring with the NSB (*r* = -.23) but is no longer present with the ASPENS (*r* = -.16). Item 3 also has a small relationship with the spring SESAT (*r* = -.22). Relations among measures besides the individual NLT Items are discussed in Research Question 1A.

The scores of the spring outcome measures were regressed on the fall NLT Items scores and fall ASPENS scores. Results are reported in Tables 11 and 12. Models including the fall ASPENS as the sole predictor are repeated from earlier for the sake of comparison. In the first model (Model 7), the spring NSB scores were regressed on the four individual fall NLT scores. In the third (Model 8), the spring NSB scores were regressed on the fall NLT Items and the fall ASPENS scores. This procedure was repeated for the remaining two models by regressing the second outcome measure, the spring SESAT scores.

Fall performance on the NLT Items explained 7% of the variance in scores on the spring NSB scores, with Item 3 being the only statistically significant predictor. Model 8, including NLT Item scores and ASPENS scores, explained 1% more variance (R^2 = .36) in the spring NSB over the ASPENS alone, $F(5, 220) = 24.62$, $p < .001$. In addition, the NLT Item 3 scores were no longer a statistically significant predictor in this combined model (p = .16).

Fall performance on the NLT Items explained 5% of the variance in scores on the spring SESAT scores, with Item 3 being the only statistically significant predictor. Model 10, including NLT Item scores and ASPENS scores, did not explain more variance (R^2 = .35) in the spring SESAT over the ASPENS alone, $F(5, 220) = 24.01$, $p < .001$. In

Table 11

Regression Results Predicting Spring NSB Performance (N = 226)

	Model 7				Model 2				Model 8			
Parameter	*b*	*SE*	*T*	*p*	*b*	*SE*	*t*	*p*	*b*	*SE*	*t*	*p*
Intercept, b_1	25.81	0.94	27.40	<.001	19.12	0.45	42.18	<.001	19.86	0.99	20.06	<.001
Fall NLT Item 1, b_2	-0.02	0.02	-1.16	.25					-0.01	0.01	-0.96	.34
Fall NLT Item 2, b_2	-0.03	0.02	-1.61	.11					0.00	0.01	-0.29	.77
Fall NLT Item 3, b_2	-0.06	0.02	-3.76	<.001					-0.02	0.01	-1.42	.16
Fall NLT Item 4, b_2	0.01	0.02	0.80	.42					0.02	0.01	1.21	.23
Fall ASPENS, b_3					0.07	0.01	10.90	<.001	0.07	0.01	9.88	<.001

Note. Model 1 R^2 = .07, *F* = 4.43, *p* < .001. Model 2 R^2 = .35, *F* = 111.80, *p* < .001. Model 3 R^2 = .36, *F* = 24.62, *p* < .001.

Table 12

Regression Results Predicting Spring SESAT Performance (N = 226)

	Model 7				Model 2				Model 8			
Parameter	*b*	*SE*	*T*	*p*	*b*	*SE*	*t*	*p*	*b*	*SE*	*t*	*p*
Intercept, b_1	31.29	1.33	23.55	<.001	22.44	0.63	35.52	<.001	22.78	1.37	16.58	<.001
Fall NLT Item 1, b_2	-0.01	0.02	-0.35	.73					0.00	0.02	0.01	.99
Fall NLT Item 2, b_2	-0.02	0.02	-0.82	.41					0.01	0.02	0.67	.50
Fall NLT Item 3, b_2	-0.07	0.02	-3.45	<.001					-0.02	0.02	-1.05	.30
Fall NLT Item 4, b_2	-0.01	0.02	-0.28	.78					0.00	0.02	-0.09	.93
Fall ASPENS, b_3					0.10	0.01	10.93	<.001	0.10	0.01	10.08	<.001

Note. Model 1 R^2 = .05, F = 3.18, p = .01. Model 2 R^2 = .35, F = 119.60, p < .001. Model 3 R^2 = .35, F = 24.01, p < .001.

addition, the NLT Item 3 scores were no longer a statistically significant predictor in this combined model (*p* = .30).

Semi-partial correlations explain the extent to which each predictor adds unique variance in explaining the outcome. Semi-partial correlations of the NLT Items and Fall ASPENS are reported in Table 13. In the combined models, the NLT Item scores add negligible explained variance (.01 or less in all cases) in the outcome measures beyond the fall ASPENS scores.

Finally, ROC analyses explored the NLT Items' abilities to correctly classify students at risk. AUC values are reported in Table 14. The NLT Items performed as well as or marginally better than chance in predicting student risk on the spring NSB (AUC = .50 to .58). Similarly, Items 1, 2, and 4 performed as well as or marginally better than chance in predicting student risk on the spring SESAT (AUC = .49 to .54). Item 3 of the NLT identified student risk on the SESAT better than chance (AUC = .64).

Table 13

Unique Variance Explained in the Outcome Measures (Semi-partial Correlations)

	Spring NSB	Spring SESAT
Fall NLT Item 1: 34	<.01	<.01
Fall NLT Item 2: 12	<.01	<.01
Fall NLT Item 3: 89	.01	<.01
Fall NLT Item 4: 57	<.01	<.01
Fall ASPENS	.28*	.30*

*$p < .01$.

Table 14

AUC for the Individual NLT Items to the Spring Outcome Measures

	Spring NSB (23rd Percentile)		Spring SESAT (25th Percentile)	
	AUC	CI	AUC	CI
NLT Item 1: 34	.55	.46-.64	.49	.41-.57
NLT Item 2: 12	.58	.49-.67	.54	.46-.62
NLT Item 3: 89	.57	.47-.67	.64	.57-.72
NLT Item 4: 57	.50	.41-.59	.52	.44-.59

IV. DISCUSSION

In the discussion, I frame the current study, offer my summarization and interpretation of the results and note limitations. Based on results and limitations of the current study, I suggest next steps and directions for future research.

Mathematics instruction in the United States consistently underserves our nation’s students, as evidenced by stagnant and unsatisfactory achievement (Center for Education Statistics, 2019). Universal screening is one mechanism for delivering critical content to the students most in need (Albers & Kettler, 2014; Shinn, 2006). However, the evidence base for early mathematics screeners is limited, thus complicating the task for schools to make accurate, useful screening decisions (Gersten et al., 2012). Due to how it integrates several key mathematical concepts, the mental number line offers promise for serving as a standalone screener or supplementing existing screening batteries (Schneider et al., 2018). The research base of the number line and its relation to mathematical development and competence derives primarily from a developmental and cognitive lens. Only two prior studies have leveraged the number line task as a screener for identifying educational risk (Clarke et al., 2018; Sutherland et al., 2020).

This work extended the exploration of the number line assessment as a screener. Due to the breadth and cumulative depth of mathematical curricula, numeracy screening benefits from assessing a range of skills (Gersten et al., 2012; Seethaler & Fuchs, 2010; Vanderheyden et al., 2017). This prompted my investigation of a number line assessment in conjunction with an established, but not maximal, multi-skill math screener. As part of a larger study, 226 kindergarten students were administered the NLT in the fall. In the fall and spring, students were also administered an established screening battery

(ASPENS) and a short outcome measure (NSB). A comprehensive outcome measure (SESAT) was administered in the spring only. Analyses explored the predictive properties of a short form four-item NLT as compared to an established screening measure. The items within the NLT were also explored for variations in decision-making utility. Particular attention was paid to practical considerations: added value and efficiency.

Summary and Interpretation of Results

The mean performance of the students in this sample does not appear to be significantly different than prior research. Number line research often reports student performance as mean absolute error or percent absolute error (Schneider et al., 2018). This study used summed absolute error (*M* = 112.27), which, averaged over four trials, gives a mean absolute error of 28.07. This mean is relatively similar to that found in the cognitive research (24% and 24%; Booth & Siegler, 2006; Siegler & Booth, 2004) and in the number line screening research (29.30 and 33.30; Clarke et al., 2018; Sutherland et al., 2020) with similar-aged students. The rest of the results should be interpreted in this context.

Research Question 1A: Association Among Early Numeracy Measures

First, relations among the early numeracy measures at various timepoints was examined. Amongst the ASPENS, NSB and SESAT at both time points, the NLT had small concurrent relations with the ASPENS and NSB in the fall (r^2 = -.26 and -.24, respectively). Other concurrent or predictive relations were insignificant and/or negligible (r^2 = -.13 to -.19). No evidence exists that the NLT possesses predictive validity to spring NSB or SESAT performance. In comparison, the ASPENS has a strong and significant concurrent association with the NSB in the fall (r^2 = .68), as well as strong and significant predictive associations with the spring NSB and SESAT (r^2 = .59 for both). The large correlations among the fall ASPENS and other numeracy measures support their tapping into similar mathematics constructs.

The small or weak associations of the NLT with other measures should not be wholly unexpected. Cognitive researchers have found that the NLT demonstrates a moderate association with general mathematical competence in elementary-aged children. Schneider et al. (2018) found that age moderated the NLT's association with general mathematical competence. For early elementary students (aged 6-9), they found an average correlation of .442 between whole-number number line estimation and mathematical competence. Prior to age 6, which likely applies to many children in the fall of kindergarten, number line estimation demonstrates an average correlation of .296 with general mathematical competencies. In the current study, with 66.4% of the sample under six years of age, the NLT demonstrated a comparable concurrent association with the ASPENS (r^2 = -.26).

Notably, the reviewed studies in Schneider et al.'s (2018) meta-analysis with participants under age 6 primarily completed number lines ranging from 0-10 and 0-20. Furthermore, results from a study by Muldoon et al. (2011) suggest that the association of number line estimation and math competence is dependent on the scale used in the task. Muldoon et al. compared 0-10, 0-20 and 0-100 number lines with Scottish and Chinese 5-year-olds and found that the Scottish children performed best on the 0-10 and 0-20 number lines. In addition, the 0-20 performance had the highest associations with other mathematical measures. Thus, low associations found in the current study may be due to the range (0-100) of the number line task used. This range was originally selected due to the mixed ranges used in prior research and due to its alignment with the numbers that kindergarten students encounter throughout the year.

Research Question 1B: Explained Variance

When considered as part of a battery with an established mathematics screener, the NLT did not add meaningful value. When either outcome measure (spring NSB or spring SESAT scores) was regressed on the two predictors (fall NLT and fall ASPENS scores), fall ASPENS performance was the only significant predictor of future performance (R^2 = .31 to .34). The NLT explained negligible and insignificant variance in the outcomes (R^2 <.01). In a prior study, Clarke and colleagues (Clarke et al., 2018) also did not find statistically significant incremental validity with either a 0-20 or 0-100 number line above the ASPENS. In a conceptual replication, Sutherland and colleagues (2020) found a 0-100 number line contributed 7% incremental validity above the ASPENS. While 7% added value holds marginal clinical significance, the literature base establishing the association between number line estimation performance with general mathematical competence and the mixed results of these screening studies suggest that task design should be explored further.

Research Question 1C: Classification Accuracy

In terms of their abilities to distinguish between students truly at risk or not at risk, the ASPENS (AUC = .83 to .86) again outperformed the NLT by a large margin (AUC = .59 for both measures). The ASPENS meets the minimum acceptable value (.75) to be effective for determining risk status (Cummings & Smolkowski, 2015). Most importantly, the 95% confidence interval for the NLT’s true AUC value includes or is very close to .50. In other words, it is very likely that this form of the NLT performs no better than chance in identifying students.

Examining the specificity and sensitivity of various cut scores on these measures enriches these conclusions. When maximizing overall classification power, the NLT and ASPENS are similarly able to identify truly "at risk" students. This is true for both the NSB (at a cut score at the 23rd percentile) and the SESAT (at a cut score at the 25th percentile). However, the ASPENS is substantially more specific than the NLT, correctly identifying 29-40% more "not at risk" students.

Schools, however, do not regard false positives and negatives equally. False positives are students who show up as "at risk" on a screener but who would be sufficiently served by their existing classroom supports. These students may be removed from the general education setting for a short, intensified intervention session multiple days a week. Resources may be spread too thin if too many students are identified as "at risk." Conversely, false negatives are students who do not show up as "at risk" on a screener but will be underserved by their environment and at risk for academic failure without the introduction of intensified supports. We also know that prevention and early intervention is cost- and time-effective. False negatives risk needing remediation, the more costly alternative to prevention.

Thus, the NLT and ASPENS were examined for their abilities to classify students while substantially reducing false negatives. Both measures are capable of catching over 90% over of the students truly "at risk". In doing so, however, the NLT correctly identified only 12-16% of the students "not at risk." The ASPENS, meanwhile, correctly identified 48-60% of the students "not at risk." Schools that would utilize the NLT for their classification decisions would either miss many students who truly need supports or would find themselves providing intensified intervention to a large swath of students who wouldn’t have been at risk receiving their existing classroom instruction.

Research Question 2: Item-Level Analyses

Looking at individual item performance on the NLT provides a lens to understanding the measure’s overall performance and potential item-level contributions. Performance on each NLT item is unexpectedly unrelated to performance on any other item. In addition, only Item 3 of the NLT is associated with the criterion math measures. This lack of internal consistency and concurrent or predictive validity suggests this number line task is not measuring related mathematical skills as expected.

In addition, the NLT items considered individually provide no evidence for incremental validity above the ASPENS in explaining the outcome measures. Interestingly, models that only include the NLT Items explain more variance in the outcomes than the full-scale NLT scores. For example, regressing the spring NSB scores on the individual NLT item scores explains more variance than regressing the spring NSB scores on the full-scale (or sum of the items) NLT scores (R^2 = .07 and .03, respectively). A similar result is seen with the spring SESAT scores (R^2 = .05 and .02, respectively). Additionally, the unique variance explained by the ASPENS is lower in models with the NLT Items (R^2 = .28 and .30) than with the full-scale NLT (R^2 = .31 and .32). These findings suggest that certain items of the NLT, compared to their peers, are more related to other math measures.

A hypothesis for the NLT Items, that Item 4 (target numeral 57) may be more informative than the others, was rejected. Rodrigues, Jordan, and Hansen (2019) found that the “simpler” items, such as 1/2 on a 0-1 fraction number line were the most predictive of general math performance. Similarly, it was hypothesized that a stimulus of 57 would be aided by a student’s abilities to a) identify the two-digit numeral correctly and b) to leverage nascent proportional reasoning to dissect the line approximately in half. It was hypothesized exhibiting these developing mathematical competencies would be associated with greater overall mathematical competence in kindergarten.

However, Item 3 (target numeral 89) was found to be most correlated with other math measures. Item 3 has small correlations with the outcome measures (r^2 = -.23 and -.22). Interestingly, the relationships between Item 3 and the outcomes are higher than the full-scale NLT scores with the outcomes (r^2 = -.19 and -.17). Comparatively, Items 1, 2, and 4 exhibit minimal relationships with the other math measures and, in the models, may simply be noise. Item 3 alone may be responsible for the full-scale’s relation to the other math measures. Thus, summing all four scores together adds noise to Item 3’s contributions.

It is speculated that the uniqueness of Item 3 is due to a large portion of the sample using an undiscerned strategy. It’s important to consider that, if students were randomly responding for all items, scores would exhibit an equal frequency of errors. Students would be just as likely to respond with 1 to a prompt to locate the number 61 on the line as they would with 100. However, the cognitive evidence base shows that, when presented with an array of similar options, people attend to and choose options near the middle most often (Atalay, Bodur, & Rasolofoarison, 2012; Christenfeld, 1995; Lo & Tsang, 2018; Rodway, Schepman, & Lambert, 2012). For children presented with unfamiliar numbers, all possible locations on the number line may seem equal. Thus, a tendency towards the center is plausible.

In examining the specific items, Items 1 (target numeral 34) and 4 (target numeral 57) are closest to the center of the number line. If students were employing the strategy hypothesized above, such that all placements on the number line are treated relatively equal, one would expect smaller average errors for items nearest the middle. In fact, post-hoc comparisons of the means revealed that students are, on average, more accurate for Items 1 and 4 than for Item 2. Thus, it is possible that a number of students were employing this pattern of responding on the NLT.

While the data is presented as absolute error, without directionality, examination of the item distributions also contributes to this theory. More so than the other items, the distribution of Item 3 is somewhat bimodal (Figure 7). These two peaks appear to be centered around 0-10 and around 40. Using these absolute errors, we can infer response patterns to some extent. With the first error peak around 0-10, we may infer one "group" of students responded in the range of 79-99. For the other peak, we can conclude directionality. Any error above 11 (due to the endpoint of 100) means the student had a negatively-oriented error, or they responded with a number less than 89. Thus, the other "group" of students responded around the midpoint of the line. Similarly, the peak of Item 2's (target numeral 12) error distribution is around 30. A cluster of errors around 40 would add credence to this theory.

This pattern of behavior is also supported by Muldoon et al.'s (2011) findings. Their data found children responded in a relatively linear pattern for numbers under 15 (on the 0-100 number line). This pattern was not observed above 15, and they concluded children were likely guessing for these numbers. It's possible students in the current study approached the task similarly and responded semi-randomly for unfamiliar numbers. However, this study had only one item below 15, limiting the ability to infer trends above or below 15.

More importantly, this theory may interact with why Item 3 is most predictive. Attention should be paid again to the bimodal nature of Item 3's distribution. This item may distinguish, more than the other items, between guessers and non-guessers. The correct placement for Items 1 (target 34) and 4 (target 57) are near the midpoint of the line, so these items may fail to distinguish students who understand these numbers and students who are using a semi-random strategy. Item 3's target numeral is the rightmost endpoint and is also a number that kindergarteners are not expected to be familiar with at school entry (Muldoon et al., 2011). In contrast, knowing (or not knowing) where to place 89 on the number line at kindergarten may be indicative of future performance (to a small degree). Future research is needed to explore item-level utility.

Implications

In an applied context, the findings from this study do not present value for schools. This study found no evidence that a four-item NLT promotes more informed screening decisions. By itself and while supplementing an established screener, this NLT does not uniquely contribute to predicting how students will perform at the end of the year. This holds true whether students are judged by a short-form outcome measure (the NSB) or a longer, comprehensive outcome measure (the SESAT). Evidence supports that, at best, educators will get a small sense of their students' mathematical competency from the NLT. However, the NLT does not provide any information that would not be better provided by the ASPENS. Schools do not have excess time to assess each student with an additional math measure, no matter how short, if it does not provide actionable information.

However, prior studies of the NLT in an educational context (Clarke et al., 2018; Sutherland et al., 2020) have found unique and meaningful contributions of the NLT in screening decisions. While it does not appear the students in this sample performed worse, on average, than other studies, a clear difference between this study and its peers is the format of the number line task employed.

In the prior cited number line screening studies, the number line assessment had 26 items compared to the four-item form in the current study. Additional items sample more behavior and, potentially, allow for a greater approximation of the construct of interest. In this case, that is general mathematical competence. The prior studies utilized random or semi-random ordering of the task items. The current study did not, limiting the ability to counteract order effects. Order effects can be substantial in number line estimation tasks as prior items can serve as a mental "anchor" that affects future placement of numbers (Siegler, 2016; White & Szucs, 2012).

These studies also utilized a descriptive task explanation and practice items with corrective feedback prior to the measure. It is possible these additions would have increased the validity of students' responses, or the likelihood that students would be able to respond in ways that reflect their true mathematical knowledge. Lastly, the current study utilized a paper and pencil format. For students in the fall of kindergarten it is unclear if motor skills would be a barrier to participation. The cited prior studies used an iPad administered form, which could also be influenced by motor skills. Future number line research should carefully consider these task differences when designing or utilizing number line estimation tasks with young children.

Rodrigues, Jordan and Hansen (2019) found that a small number of items on their number line measure held a disproportionate amount of the predictive power. The items most predictive of future performance were the “simpler” items such as 1/2 or 5/6 on a fraction number line. One would assume the simplest item on this shortened NLT would be Item 2 (target numeral 12) but this was not the case. Items 1 (target number 34) and 4 (target number 57) appeared to be the easiest. One theory is that this range of numbers may have been inappropriate for this age range. However, as mentioned earlier, prior studies have utilized the 0-100 number line to some success. These prior studies demonstrated similar amounts of error by participants and yet derived greater value from their number line measures. The task may have been misunderstood by participants, due to limited directions and practice items.

A goal of this study was to explore whether a measure that optimizes efficiency (by reducing the number of items and thus the time needed to administer) could maintain its predictive value demonstrated in prior studies. This study does not support that this four-item form of number line estimation succeeds in this mission. However, the pursuit of a screener that efficiently leverages a smaller selection of highly predictive items still holds potential value for school practice.

Limitations and Future Research

Interpreting these findings should be made in the context of the study’s limitations. While linear regression is relatively robust to violations of assumptions, the data in this study violated numerous assumptions. The most striking of these violations is the bimodal nature of the NLT summed scores, which is also seen in the distribution of Item 3. The ASPENS in the fall also violates the assumption of normality, though the distribution more closely resembles a normal distribution in the spring. It is possible that kindergarteners who are new to formal education more closely resemble a normal distribution after a year of instruction. Regardless, these violations impair the confidence of the results.

Another limitation is the relatively small sample size, which limits the extent to which this sample could generalize to a typical school. Furthermore, 25 students were excluded who did not provide an appropriate response to all four NLT items. In other words, over 10% of the potential sample was removed. This may include both students who misunderstood the task or were not attending. Representing these students in the sample may have improved the predictive value of the NLT.

The construction of the number line task presents a number of limitations and future directions. Firstly, the NLT had a response range for each item of 0-100. Unlike the other measures used, which dichotomize responses as right or wrong, the NLT allows for a continuous spectrum of responses. This can create situations where, given a stimulus of 12, one student responded with position 65 on the number line and another responded with position 86. Is the former student demonstrating greater understanding of the numeral 12? This point is particularly salient considering prior studies tend to administer a 0-10 or 0-20 number line to young children (Schneider et al., 2018) or found greater associations with a 0-20 number line than a 0-100 number line task (Muldoon et al., 2011). Alternatively, number lines with smaller ranges may be more closely linked to students’ developmental level at kindergarten entry. Due to the mixed evidence regarding range for the number line estimation task, future research should explore varying number line ranges for predicting future performance and different methods for quantifying student responses.

The background information and directions provided to participating students prior to the NLT was limited, especially in comparison to other studies. Given the novel nature of the task compounded with kindergarten students' newness to formal education, more detailed directions including teaching items may be critical to ensure that students fully understand task demands. Future studies are urged to explain the task to participants and provide practice items to increase the possibility of measuring a participant's true mathematical competence.

Studies demonstrating promise of the 0-100 NLT with young children have utilized a 26-item form, unlike the short four-item form used here. Additional items, or additional behaviors sampled, appear to increase the NLT's association with future math achievement in young children. However, the balance between efficiency and a minimal necessary amount of items is unanswered. This point is especially salient knowing that the administration time of the 26-item form utilized by Sutherland and colleagues (2020) commonly met the self-imposed five-minute limit.

In the pursuit of consolidated measures, the specific items selected for stimuli should be considered. Because of the current finding that certain items may have added noise to the model and detracted from the value of the full-scale, identifying high-value items may be critical for increasing the value of the NLT as a screener. As opposed to randomly sampling across the chosen number range, items could be strategically selected based on prior hypotheses or data.

Because of the evidence base supporting the logarithmic to linear shift in young children, oversampling 0-20 on a 0-100 number line may be more developmentally appropriate. Additionally, the Common Core calls for kindergarteners to be able to count by both ones and tens to 100 (Practices, 2010). Oversampling the decades may align with and be sensitive to students' response to classroom instruction. Older students are observed to rely upon familiar anchor points (such as 25, 50, or 75; D. Cohen & Sarnecka, 2016). While kindergarteners may not use these numbers as benchmarks, they may use numbers within their familiar range, like 5, 10, and 15 instead.

Research including such items can illuminate which are most predictive. However, the consolidation of items should be performed post-hoc, once the effectiveness of particular items is established. Then, screeners can maximize efficiency by including only these high-utility items. Finally, counterbalancing these items would reduce the influence of potential order effects.

Future number line research is urged to explore various task forms to design more effective and efficient screeners. Certain factors should be consistent, such as task directions and practice items, while other factors would benefit from variation across forms, times, and skill levels.

Conclusion

Schools need accurate and efficient screeners that empower them to make better decisions around mathematics instruction and intervention. While this study does not provide evidence of this form of the NLT filling that gap, other studies have demonstrated unique contributions of assessing the mental number line. Evidence supports the number line's association with mathematics concepts spanning diverse competencies and ages, highlighting the potential for a versatile mathematics screener across school grades. However, practical implementation is crucial. Though schools exist in a context that demands efficiency and accounting for every minute, this study cautions that maximizing efficiency may sacrifice clinical utility. Knowing the long-reaching implications of successful early prevention and intervention, future research should strive to increase schools' abilities to make informed decisions around who to serve.

REFERENCES CITED

Albers, C. A., & Kettler, R. J. (2014). Best practices in universal screening. *Best Practices in School Psychology: Data-Based and Collaborative Decision Making*, 121–131.

Ashcraft, M. H., & Moore, A. M. (2012). Cognitive processes of numerical estimation in children. *Journal of Experimental Child Psychology*, *111*(2), 246–267. https://doi.org/10.1016/j.jecp.2011.08.005

Atalay, A. S., Bodur, H. O., & Rasolofoarison, D. (2012). Shining in the center: Central gaze cascade effect on product choice. *Journal of Consumer Research*, *39*(4), 848–866. https://doi.org/10.1086/665984

Baker, S., Gersten, R., Flojo, J., Katz, R., Chard, D., & Clarke, B. (2002). *Preventing mathematics difficulties in young children: Focus on effective screening of early number sense delays*. Technical Report.

Balu, R., Zhu, P., Doolittle, F., Schiller, E., Jenkins, J., & Gersten, R. (2015). Evaluation of response to intervention practices for elementary school reading. NCEE 2016-4000. *National Center for Education Evaluation and Regional Assistance*.

Baroody, A. J. (2002). The developmental foundations of number and operation sense. In *Poster presented at the EHR/REC (NSF) Principal Investigators' Meeting ("Learning and Education: Building Knowledge, Understanding Its Implications"), Arlington, VA*.

Barth, H. C., & Paladino, A. M. (2011). The development of numerical estimation: Evidence against a representational shift. *Developmental Science*, *14*(1), 125–135. https://doi.org/10.1111/j.1467-7687.2010.00962.x

Berch, D. B. (2005). Making sense of number sense: Implications for children with mathematical disabilities. *Journal of Learning Disabilities*, *38*(4), 333–339. https://doi.org/10.1177/00222194050380040901

Berteletti, I., Lucangeli, D., Piazza, M., Dehaene, S., & Zorzi, M. (2010). Numerical estimation in preschoolers. *Developmental Psychology*, *46*(2), 545–551. https://doi.org/10.1037/a0017887

Bodovski, K., & Farkas, G. (2007). Do instructional practices contribute to inequality in achievement? The case of mathematics instruction in kindergarten. *Journal of Early Childhood Research*, *5*(3), 301–322.

Booth, J. L., & Siegler, R. (2006). Developmental and individual differences in pure numerical estimation. *Developmental Psychology*, *42*(1), 189–201. https://doi.org/10.1037/0012-1649.41.6.189

Boyer, T. W., Levine, S. C., & Huttenlocher, J. (2008). Development of proportional reasoning: Where young children go wrong. *Developmental Psychology*, *44*(5), 1478–1490. https://doi.org/10.1037/a0013110

Case, R. (1998). A psychological model of number sense and its development. In *annual meeting of the American Educational Research Association, San Diego*.

Case, R., Okamoto, Y., Griffin, S., McKeough, A., Bleiker, C., Henderson, B., … Keating, D. P. (1996). The role of central conceptual structures in the development of children’s thought. *Monographs of the Society for Research in Child Development*, i–295.

Center for Education Statistics, N. (2015). 2015 NAEP mathematics grades 4 and 8 assessment report cards: Summary data tables for national and state average scores and achievement level results. Retrieved from https://www.nationsreportcard.gov/reading_math_2015/files/2015_Results_Appendix_Math.pdf

Chan, C., Chan, G. C. H., Leeper, T. J., & Becker, J. (2018). rio: A Swiss-army knife for data file I/O.

Chard, D. J., Clarke, B., Baker, S., Otterstedt, J., Braun, D., & Katz, R. (2005). Using measures of number sense to screen for difficulties in mathematics: Preliminary findings. *Assessment for Effective Intervention*, *30*(2), 3–14. https://doi.org/10.1177/073724770503000202

Christenfeld, N. (1995). Choices from identical options. *Psychological Science*, *6*(1), 50–55. https://doi.org/10.1111/j.1467-9280.1995.tb00304.x

Clarke, B., Baker, S., Smolkowski, K., & Chard, D. J. (2008). An analysis of early numeracy curriculum-based measurement: Examining the role of growth in student outcomes. *Remedial and Special Education*, *29*(1), 46–57. https://doi.org/10.1177/0741932507309694

Clarke, B., Doabler, C. T., Fien, H., Baker, S. K., & Smolkowski, K. (2012). A randomized control trial of a Tier 2 kindergarten mathematics intervention (Project ROOTS). US Department of Education, Institute of Education Sciences. *Special Education Research, CFDA*, (84.324), 2012–2016.

Clarke, B., Gersten, R. M., Dimino, J., & Rolfhus, E. (2011). Assessing student proficiency of number sense (ASPENS). *Longmont, CO: Cambium Learning Group, Sopris Learning*.

Clarke, B., & Shinn, M. R. (2004). A preliminary investigation into the identification and development of early mathematics curriculum-based measurement. *School Psychology Review*, *33*(2), 234–248.

Clarke, B., Strand Cary, M. G., Shanley, L., & Sutherland, M. (2018). Exploring the promise of a number line assessment to help identify students at-risk in mathematics. *Assessment for Effective Intervention*, 153450841879173. https://doi.org/10.1177/1534508418791738

Clements, D. H. (1999). Subitizing: What is it? Why teach it? *Teaching Children Mathematics*, *5*, 400–405.

Clements, D. H., Sarama, J., & DiBiase, A.-M. (2003). *Engaging young children in mathematics: Standards for early childhood mathematics education*. Routledge.

Cohen, D., & Sarnecka, B. W. (2016). Children’s number-line estimation shows development of measurement skills (not number representations). *Developmental Psychology*, *93*(4), 292–297. https://doi.org/10.1016/j.contraception.2015.12.017.Women

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155.

Conoyer, S. J., Foegen, A., & Lembke, E. S. (2016). Early numeracy indicators: Examining predictive utility across years and states. https://doi.org/10.1177/0741932515619758

Cooper Jr, R. G. (1984). Early number development. In *Origins of cognitive skills: The eighteenth annual Carnegie symposium on cognition* (pp. 157–192). Erlbaum.

Cross, C. T., Woods, T. A., & Schweingruber, H. E. (2009). *Mathematics learning in early childhood: Paths toward excellence and equity*. National Academies Press. https://doi.org/10.17226/12519

Cummings, K. D., & Smolkowski, K. (2015). Selecting students at risk of academic difficulties. *Assessment for Effective Intervention*, *41*(1), 55–61. https://doi.org/10.1177/1534508415590396

Curtin, J. (2018). lmSupport: Support for linear models. Retrieved from https://cran.r-project.org/package=lmSupport

Dehaene, S. (2001). Précis of the number sense. *Mind and Language*, *16*(1), 16–36. https://doi.org/10.1111/1468-0017.00154

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude access to parity and magnitude knowledge during number processing. *Journal of Experimental Psychology: General*, *122*(3), 371–396.

Dehaene, S., Dupoux, E., & Mehler, J. (1990). Is numerical comparison digital? Analogical and symbolic effects in two-digit number comparison. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(3), 626.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., … Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, *43*(6), 1428–1446. https://doi.org/10.1037/0012-1649.43.6.1428

Feigenson, L., Libertus, M. E., & Halberda, J. (2013). Links between the intuitive sense of number and formal mathematics ability. *Child Development Perspectives*, *7*(2), 74–79. https://doi.org/10.1111/cdep.12019

Fletcher, J. M., & Vaughn, S. (2009). Response to intervention: Preventing and remediating academic difficulties. *Child Development Perspectives*, *3*(1), 30–37. https://doi.org/10.1111/j.1750-8606.2008.00072.x

Foegen, A., Jiban, C., & Deno, S. (2007). Progress monitoring measures in mathematics: A review of the literature. *The Journal of Special Education*, *41*(2), 121–139.

Friso-van den Bos, I., Kroesbergen, E. H., Van Luit, J. E. H., Xenidou-Dervou, I., Jonkman, L. M., Van der Schoot, M., & Van Lieshout, E. C. D. M. (2015). Longitudinal development of number line estimation and mathematics performance in primary school children. *Journal of Experimental Child Psychology*, *134*, 12–29. https://doi.org/10.1016/j.jecp.2015.02.002

Fuchs, D., & Fuchs, L. S. (2017). Critique of the national evaluation of response to intervention: A case for simpler frameworks. *Exceptional Children*, *83*(3), 255–268. https://doi.org/10.1177/0014402917693580

Fuchs, L. S., Fuchs, D., Compton, D. L., Bryant, J. D., Hamlett, C. L., & Seethaler, P. M. (2007). Mathematics screening and progress monitoring at first grade: Implications for responsiveness to intervention. *Exceptional Children*, *73*(3), 311–330. https://doi.org/10.1177/001440290707300303

Fuchs, L. S., Fuchs, D., Hamlett, C. L., Thompson, A., Roberts, P. H., Kubek, P., & Stecker, P. M. (1994). Technical features of a mathematics concepts and applications curriculum-based measurement system. *Diagnostique*, *19*(4), 23–49.

Gallistel, C. R., & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition*, *44*(1–2), 43–74.

Geary, D. C. (2011). Consequences, characteristics, and causes of mathematical learning disabilties and persistent low achievement in mathematics. *Journal of Devleopmental Behaviour Pediatrics*, *32*(3), 250–263. https://doi.org/10.1097/DBP.0b013e318209edef.Consequences

Geary, D. C., Hoard, M. K., Nugent, L., & Byrd-Craven, J. (2008). Development of number line representations in children with mathematical learning disability. *Developmental Neuropsychology*, *33*(3), 277–299. https://doi.org/10.1080/87565640801982361

Gersten, R., Beckmann, S., Clarke, B., Foegen, A., Marsh, L., Star, J. R., & Witzel, B. (2009). Assisting students struggling with mathematics: Response to Intervention (RtI) for elementary and middle schools. *What Works Clearinghouse*. https://doi.org/10.1016/j.jhazmat.2011.04.026

Gersten, R., Clarke, B., Jordan, N. C., Newman-Gonchar, R., Haymond, K., & Wilkins, C. (2012). Universal screening in mathematics for the primary grades: Beginnings of a research base. *Exceptional Children*, *78*(4), 423–445. https://doi.org/10.1177/001440291207800403

Gersten, R., Jordan, N. C., & Flojo, J. R. (2005). Early identification and interventions for students with mathematics difficulties. *Journal of Learning Disabilities*, *38*(4), 293–304. https://doi.org/10.1177/00222194050380040301

Goode, K., & Rey, K. (2019). ggResidpanel: Panels and interactive versions of diagnostic plots using "ggplot2." Retrieved from https://cran.r-project.org/package=ggResidpanel

Griffin, S. (2002). The development of math competence in the preschool and early school years: Cognitive foundations and instructional strategies. *Mathematical Cognition*, 1–32.

Griffin, S. (2004). Building number sense with Number Worlds: A mathematics program for young children. *Early Childhood Research Quarterly*, *19*(1), 173–180.

Griffin, S., Case, R., & Siegler, R. S. (1994). *Rightstart: Providing the central conceptual prerequisites for first formal learning of arithmetic to students at risk for school failure.* The MIT Press.

Hampton, D. D., Lembke, E. S., Lee, Y. S., Pappas, S., Chiong, C., & Ginsburg, H. P. (2012). Technical adequacy of early numeracy curriculum-based progress monitoring measures for kindergarten and first-grade students. *Assessment for Effective Intervention*, *37*(2), 118–126. https://doi.org/10.1177/1534508411414151

Hansen, N. (2015). *Development of fraction knowledge: A longitudinal study from third through sixth grade (Doctoral dissertation)*. University of Delaware.

Hansen, N., Jordan, N. C., & Rodrigues, J. (2017). Identifying learning difficulties with fractions: A longitudinal study of student growth from third through sixth grade. *Contemporary Educational Psychology*, *50*, 45–59. https://doi.org/10.1016/j.cedpsych.2015.11.002

Harcourt Educational Measurement. (2002). *Stanford Achievement Test-Tenth edition*. San Antonio, Texas: Author.

Harrell Jr, F. E. (2020). Hmisc: Harrell Miscellaneous. Retrieved from https://cran.r-project.org/package=Hmisc

Hiebert, J., & Wearne, D. (1996). Instruction, understanding, and skill in multidigit addition and subtraction. *Cognition and Instruction*, *14*(3), 251–283. https://doi.org/10.1207/s1532690xci1403_1

Hudson, P., & Miller, S. P. (2005). *Designing and implementing mathematics instruction for students with diverse learning needs*. Allyn & Bacon.

Individuals with Disabilities Education Act (2004). 20 U.S.C. § 1400.

Jordan, N. C., Glutting, J., & Ramineni, C. (2008). *A number sense assessment tool for identifying children at risk for mathematical difficulties*. Elsevier Inc. https://doi.org/10.1016/B978-012373629-1.50005-8

Jordan, N. C., Glutting, J., & Ramineni, C. (2010). The importance of number sense to mathematics achievement in first and third grades. *Learning and Individual Differences*, *20*(2), 82–88. https://doi.org/10.1016/j.lindif.2009.07.004

Jordan, N. C., Glutting, J., Ramineni, C., & Watkins, M. W. (2010). Validating a number sense screening tool for use in kindergarten and first grade: Prediction of mathematics proficiency in third grade. *School Psychology Review*, *39*(2), 181–195.

Jordan, N. C., Hansen, N., Fuchs, L. S., Siegler, R. S., Gersten, R., & Micklos, D. (2013). Developmental predictors of fraction concepts and procedures. *Journal of Experimental Child Psychology*, *116*(1), 45–58. https://doi.org/10.1016/j.jecp.2013.02.001

Jordan, N. C., Kaplan, D., & Hanich, L. B. (2002). Achievement growth in children with learning difficulties in mathematics: Findings of a two-year longitudinal study. *Journal of Educational Psychology*, *94*(3), 586–597. https://doi.org/10.1037//0022-0663.94.3.586

Jordan, N. C., Kaplan, D., Olah, L. N., & Locuniak, M. N. (2006). Number sense growth in kindergarten: A longitudinal investigation of children at-risk for mathematics difficulities. *Child Development*, *77*(1), 153–177. https://doi.org/10.1111/j.1467-8624.2006.00862.x

Jordan, N. C., Kaplan, D., Ramineni, C., & Locuniak, M. N. (2009). Early math matters: Kindergarten number competence and later mathematics outcomes. *Developmental Psychology*, *45*(3), 850–867. https://doi.org/10.1037/a0014939

Judge, S., & Watson, S. M. R. (2011). Longitudinal outcomes for mathematics achievement for students with learning disabilities. *Journal of Educational Research*, *104*(3), 147–157. https://doi.org/10.1080/00220671003636729

Laski, E. V., Casey, B. M., Yu, Q., Dulaney, A., Heyman, M., & Dearing, E. (2013). Spatial skills as a predictor of first grade girls' use of higher level arithmetic strategies. *Learning and Individual Differences*, *23*, 123–130. https://doi.org/10.1016/J.LINDIF.2012.08.001

Laski, E. V., & Siegler, R. (2007). Is 27 a big number? correlational and causal connections among numerical categorization, number line estimation, and numerical magnitude comparison. *Child Development*, *78*(6), 1723–1743. https://doi.org/10.1111/j.1467-8624.2007.01087.x

Lee, Y. S., & Lembke, E. (2016). Developing and evaluating a kindergarten to third grade CBM mathematics assessment. *ZDM - Mathematics Education*, *48*(7), 1019–1030. https://doi.org/10.1007/s11858-016-0788-6

Lembke, E., & Foegen, A. (2009). Identifying early numeracy indicators for kindergarten and first-grade students. *Learning Disabilities Research & Practice, 24*(1), 12–20. https://doi.org/10.1111/j.1540-5826.2008.01273.x

Lo, L. Y., & Tsang, C. Y. (2018). Best thing is always in the middle? An investigation of centrality preference by eye-tracking technique and memory recall. *Journal of Pacific Rim Psychology*, *12*. https://doi.org/10.1017/prp.2018.5

Mazzocco, M. M. M. (2005). Challenges in identifying target skills for math disability screening and intervention. *Journal of Learning Disabilities*, *38*(4), 318–323. https://doi.org/10.1177/00222194050380040701

Mazzocco, M. M. M., & Thompson, R. E. (2005). Kindergarten predictors of math learning disability. *Learning Disabilities Research and Practice*, *20*(3), 142–155. https://doi.org/10.1111/j.1540-5826.2005.00129.x

Mcclure, E. R., Guernsey, L., Clements, D. H., Bales, S. N., Nichols, J., Kendall-Taylor, N., & Levine, M. H. (2017). STEM starts early: Grounding science, technology, engineering, and math education in early childhood. Retrieved from http://joanganzcooneycenter.org/publication/stem-starts-early/

Mix, K. S., Huttenlocher, J., & Levine, S. C. (2002). *Quantitative development in infancy and early childhood*. Oxford University Press.

Morgan, P. L., Farkas, G., & Wu, Q. (2009). Five-year growth trajectories of kindergarten children with learning difficulties in mathematics. *Journal of Learning Disabilities*, *42*(4), 306–321. https://doi.org/10.1177/0022219408331037

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature*, *215*(5109), 1519–1520.

Muldoon, K., Simms, V., Towse, J., Menzies, V., & Yue, G. (2011). Cross-cultural comparisons of 5-year-olds' estimating and mathematical ability. *Journal of Cross-Cultural Psychology*, *42*(4), 669–681. https://doi.org/10.1177/0022022111406035

Müller, K. (2017). here: A simpler way to find your files. Retrieved from https://cran.r-project.org/package=here

National Center for Education Statistics. (2015). Postsecondary attainment: Differences by socioeconomic status. *The Condition of Education*, 1–7. Retrieved from https://nces.ed.gov/programs/coe/pdf/coe_tva.pdf

National Mathematics Advisory Panel. (2008). *Foundation for success: The final report of the national mathematics advisory panel. U.S. Department of Education* (Vol. 37). Washington, DC. https://doi.org/10.3102/0013189X08329195

National Research Council. (2001). *Adding it up: Helping children learn mathematics*. Washington, D.C.

National Science Board. (2015). Revisiting the STEM workforce: A companion to science and engineering indicators 2014. *Arlington*. National Science Foundation VA.

OECD, O. (2012). *Equity and quality in education: Supporting disadvantaged students and schools*. *Computer-Supported Collaborative Learning Conference, CSCL*. OECD Publishing Paris.

Okamoto, Y., Case, R., & Maes. (1996). Exploring the microstructure of children's central conceptual structures in the domain of number. *Monographs of the Society for Research in Child Development*, *61*(1-2), 27–58.

Olson, J. F., Martin, M. O., & Mullis, I. V. S. (2008). *TIMSS 2007 technical report*. TIMSS & PIRLS International Study Center.

Pedhazur, E. J., & Kerlinger, F. N. (1982). *Multiple regression in behavioral research*. Holt, Rinehart, and Winston.

Peeters, D., Degrande, T., Ebersbach, M., Verschaffel, L., & Luwel, K. (2016). Children's use of number line estimation strategies. *European Journal of Psychology of Education*, *31*(2), 117–134. https://doi.org/10.1007/s10212-015-0251-z

Phillips, G. W. (2007). *Chance favors the prepared mind: Mathematics and science indicators for comparing states and nations*. *American Institutes for Research*. Washington, DC. https://doi.org/10.1097/CCM.0b013e31820e6be4

Practices, N. G. A. C. for B. (2010). *Common Core State Standards for Mathematics*. *Common Core State Standards Initiative*.

Purpura, D. J., Reid, E. E., Eiland, M. D., & Baroody, A. J. (2015). Using a brief preschool early numeracy skills screener to identify young children with mathematics difficulties. *School Psychology Review*, *44*(1), 41–59. https://doi.org/10.17105/SPR44-1.41-59

R Development Core Team, R. (2011). *R: A language and environment for statistical computing*. *R Foundation for Statistical Computing*. https://doi.org/10.1007/978-3-540-74686-7

Ritchie, S. J., & Bates, T. C. (2013). Enduring links from childhood mathematics and reading achievement to adult socioeconomic status. *Psychological Science*, *24*(7), 1301–1308. https://doi.org/10.1177/0956797612466268

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77.

Rodrigues, J., Jordan, N. C., & Hansen, N. (2019). Identifying fraction measures as screeners of mathematics risk status. *Journal of Learning Disabilities*, *52*(6), 480–497. https://doi.org/10.1177/0022219419879684

Rodway, P., Schepman, A., & Lambert, J. (2012). Evidence for the centre-stage effect, *222*(July 2011), 215–222.

Schneider, M., Grabner, R. H., & Paetsch, J. (2009). Mental number line, number line estimation, and mathematical achievement: Their interrelations in grades 5 and 6. *Journal of Educational Psychology*, *101*(2), 359–372. https://doi.org/10.1037/a0013840

Schneider, M., Merz, S., Stricker, J., De Smedt, B., Torbeyns, J., Verschaffel, L., & Luwel, K. (2018). Associations of number line estimation with mathematical competence: A meta-analysis. *Child Development*, *89*(5), 1467–1484. https://doi.org/10.1111/cdev.13068

Schulte, A. C., & Stevens, J. J. (2015). Once, sometimes, or always in special education: Mathematics growth and achievement gaps. *Exceptional Children*, *81*(3), 370–387. https://doi.org/10.1177/0014402914563695

Seethaler, P. M., & Fuchs, L. S. (2010). The predictive utility of kindergarten screening for math difficulty. *Exceptional Children*, *77*(1), 37–59. https://doi.org/10.1177/001440291007700102

Shinn, M. R. (2006). Best practices in using curriculum-based measurement in a problem-solving model. *Best Practices in School Psychology V*, *I*(c), 243–261. Retrieved from http://www.nasponline.org/publications/booksproducts/bp5.aspx

Siegler, R. (2016). Magnitude knowledge: The common core of numerical development. *Developmental Science*, *19*(3), 341–361. https://doi.org/10.1111/desc.12395

Siegler, R., & Booth, J. (2004). Development of numerical estimation in young children. *Child Development*, *75*(2), 428–444. https://doi.org/10.1111/j.1467-8624.2004.00684.x

Siegler, R., & Lortie-Forgues, H. (2014). An integrative theory of numerical development. *Child Development Perspectives*, *8*(3), 144–150. https://doi.org/10.1111/cdep.12077

Siegler, R., & Opfer, J. E. (2003). The development of numerical estimation: Evidence for multiple representations of numerical quantity. *Psychological Science*, *14*(3), 237–243. https://doi.org/10.1111/1467-9280.02438

Siegler, R., Thompson, C. A., & Opfer, J. E. (2009). The logarithmic-to-linear shift: One learning sequence, many tasks, many time scales. *Mind, Brain, and Education*, *3*(3), 143–150. https://doi.org/10.1111/j.1751-228X.2009.01064.x

Siegler, R., Thompson, C. A., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive Psychology*, *62*(4), 273–296. https://doi.org/10.1016/j.cogpsych.2011.03.001

Simmons, D. C., Kame'enui, E. J., Good, R. H., Harn, B., Cole, C., & Braun, D. (2000). Building, implementing, and sustaining a beginning reading model: School by school and lessons learned. *OSSC Bulletin*, *43*(3), 3–30.

Starkey, P., & Cooper, R. G. (1980). Perception of numbers by human infants. *Science*, *210*(4473), 1033–1035.

Starkey, P., Spelke, E. S., & Gelman, R. (1990). Numerical abstraction by human infants. *Cognition*, *36*(2), 97–127. https://doi.org/10.1016/0010-0277(90)90001-Z

Starr, A., Libertus, M. E., & Brannon, E. M. (2013). Number sense in infancy predicts mathematical abilities in childhood. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(45), 18116–18120. https://doi.org/10.1073/pnas.1302751110

Sutherland, M., Clarke, B., Nese, J. F. T., Strand Cary, M., Shanley, L., Furjanic, D., & Durán, L. (2020). Investigating the utility of a kindergarten number line assessment compared to an early numeracy screening battery.

Torgesen, J. K. (2000). Individual differences in response to early interventions in reading: The lingering problem of treatment resisters. *Learning Disabilities Research & Practice*, *15*(1), 55–64.

Torgesen, J. K. (2002). The orevention of reading difficulties. *Journal of School Psychology*, *40*(1), 7–26.

Torgesen, J. K., Alexander, A. W., Wagner, R. K., Rashotte, C. A., Voeller, K. K. S., & Conway, T. (2001). Intensive remedial instruction for children with severe reading disabilities: Immediate and long-term outcomes from two instructional approaches. *Journal of Learning Disabilities*, *34*(1), 33–58. Retrieved from http://hdl.handle.net/10829/5385

Vanderheyden, A. M., Codding, R., & Martin, R. (2017). Relative value of common screening measures in mathematics. *School Psychology Review*, *46*(1), 65–87. https://doi.org/10.17105/SPR46-1.65-87

VanDerHeyden, A. M., Witt, J. C., Naquin, G., & Noell, G. (2001). The reliability and validity of curriculum-based measurement readiness probes for kindergarten students. *School Psychology Review*, *30*(3), 363–382.

Vaughn, S., & Wanzek, J. (2014). Intensive interventions in reading for students with reading disabilities: Meaningful impacts. *Learning Disabilities Research and Practice*, *29*(2), 46–53. https://doi.org/10.1111/ldrp.12031

Vaughn, S., Wexler, J., Roberts, G., Barth, A. A., Cirino, P. T., Romain, M. A., … Denton, C. A. (2011). Effects of individualized and standardized interventions on middle school students with reading disabilities. *Exceptional Children*, *77*(4), 391–407. https://doi.org/10.1177/001440291107700401

Wagner, S. H., & Walters, J. (1982). A longitudinal analysis of early number concepts: From numbers to number. *Action and Thought: From Sensorimotor Schemes to Symbolic Operations*, 137–161.

Walker, H. M., Horner, R. H., Sugai, G., Bullis, M., Sprague, J. R., Bricker, D., & Kaufman, M. J. (1996). Integrated approaches to preventing antisocial behavior patterns among school-age children and youth. *Journal of Emotional and Behavioral Disorders*, *4*(4), 194–209.

Watts, T. W., Duncan, G. J., Siegler, R. S., & Davis-Kean, P. E. (2014). What's past is prologue: Relations between early mathematics knowledge and high school achievement. *Educational Researcher*, *43*(7), 352–360. https://doi.org/10.3102/0013189X14553660

Wei, X., Lenz, K. B., & Blackorby, J. (2013). Math growth trajectories of students With disabilities: Disability category, gender, racial, and socioeconomic status differences from ages 7 to 17. *Remedial and Special Education*, *34*(3), 154–165. https://doi.org/10.1177/0741932512448253

White, S. L. J. J., & Szucs, D. (2012). Representational change and strategy use in children ' s number line estimation during the first years of primary school. *Behavioral and Brain Functions*, *8*, 1–12. https://doi.org/10.1186/1744-9081-8-1

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. (2019). Welcome to the {tidyverse}. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., & Miller, E. (2019). haven: Import and export "SPSS", "Stata" and "SAS" files. Retrieved from https://cran.r-project.org/package=haven

Wilke, C. O. (2019). cowplot: Streamlined plot theme and plot annotations for "ggplot2." Retrieved from https://cran.r-project.org/package=cowplot

Wood, G., Willmes, K., Nuerk, H.-C., & Fischer, M. H. (2008). On the cognitive link between space and number: a meta-analysis of the SNARC effect. *Psychology Science*.

Wu, H. (2013). ggROC: package for roc curve plot with ggplot2. Retrieved from https://cran.r-project.org/package=ggROC

Wynn, K. (1992). Addition and subtraction by human infants. *Nature*, *358*(6389), 749–750.

Wynn, K., Bloom, P., & Chiang, W.-C. (2002). Enumeration of collective entities by 5-month-old infants. *Cognition*, *83*(3), B55–B62.